EVALUATION OF THE EFFECTIVENESS AND ACCURACY OF PREDICTIVE MODELING TECHNIQUES ON HEALTHCARE CLAIMS IN NIGERIA

Ajijola Lukman Abolaji[a], Akindipe Oluwaleke Ebenezer[b], Lawal Abdulrasheed Olajide[c], Afuwape Omolade Moses[d], Olaleye Priscilla Oluwaseyitan[e]

Corresponding author: [a]*lajijola@unilag.edu.ng*

ABSTRACT

Predictive modeling in healthcare is essential for effective financial planning and strategic decision-making in the ever-revolving landscape of healthcare. This study, evaluate the effectiveness and accuracy of predictive modelling techniques on healthcare claims in Nigeria. Exploratory research design was used to carry out this study. The data were sourced from a reputable Health Maintenance Organization (HMO) in Nigeria with Datasets titled health datasets which have four (4) numerical features (claims paid, claim service type, categories of policyholder, gender). Descriptive statistics, testing for the performance of the model and model prediction were done on the data by using regression analysis. The paper emphasizes how predictive modeling methods might improve healthcare cost forecasts and decision-making. Results highlighted the necessity of better claims processing technologies, data analytics, and policy frameworks to increase Nigerian healthcare availability and affordability. The significant relationships between claim service types, categories of policyholder, Gender, and claims paid provide valuable information about the dynamics of healthcare claims in Nigeria.

Keywords*:* Claim, Expenditure, Healthcare, Modelling, Prediction

1. INTRODUCTION

Different countries are plague with different with health system problems globally (Oleribe, Momoh, Uzochukwu, Mbofana, Adebiyi, Barbera, Williams and Taylor-Robinson, 2019). While health service delivery challenges are more often seen in countries with a very high Human Development Index (HDI), human resources challenges attract more attention within those with a low HDI (Roncarolo, Boivin, Denis, Hébert and Lehoux, 2017). Healthcare systems in Africa have, over the years, suffered from man-made issues which cut across institutional, human resources, financial, technical and political developments. The financial aspects of healthcare are a significant concern in industrialized economies, with health insurance claims experiencing a constant increase over the past decade. This trend is evident in countries like the United States, UK, Germany, France, and Asia, where the older population is the target demographic. In developing nations, such as Brazil, India, Zimbabwe, and Sub-Saharan Africa, health insurance claims expenditures show unique dynamics due to infectious diseases, socio-economic challenges, and lifestyle factors.

The Nigerian healthcare system has faced numerous challenges since gaining independence in 1960, including limited funding, inadequate infrastructure, and poor access to quality healthcare services. The World Health Organization (2012) states that funding healthcare involves generating, conserving, and distributing funds to meet people's individual and group health requirements within the healthcare system. The rising healthcare costs in Nigeria are driven by the country's rapidly growing population, high prevalence of diseases, economic challenges, insufficient healthcare infrastructure, administrative inefficiencies, and low levels of insurance coverage. Low funding from various government tiers has led to inadequate health infrastructural facilities and poor access to quality healthcare services. Out-of-pocket (OOP) health spending, which accounts for 95% of private expenditure on health, limits the ability of poor households to access and utilize basic healthcare services. This high level of OOP health spending can create a chain reaction that influences healthcare delivery and costs, making it difficult for Nigeria's healthcare system and its health insurance sector to offer equitable, affordable, and high-quality care to its citizens, hence the need for accurate prediction of healthcare cost.

Predictive modeling in healthcare is essential for effective financial planning and strategic decision-making in the ever-revolving landscape of healthcare (Ajijola, Ojikutu and Adeleke,

2018a). Accurate prediction of healthcare expenses leads to a more effective and long-lasting healthcare insurance system, benefiting policyholders, providers, and patients in equal measure by encouraging improved budgetary management, resource distribution, and care quality. With the increasing global healthcare expenses, there is an urgent need for strong predictive modeling techniques to anticipate future costs. In contemporary times, actuarial modeling of insurance claims has become an essential research area in the health insurance sector, mainly applied in setting effective premiums. By harnessing historical data on patient demographics, medical history, clinical encounters, treatment modalities, and associated costs in Nigeria, predictive models can generate insights into future healthcare expenditures at various levels, including individual patient encounters, population segments, healthcare facilities, and entire healthcare systems. Machine learning algorithms have also proven to yield accurate results in predicting high-cost, high-need patient expenditures, so insurance companies are increasingly turning to machine learning approaches to improve their policies and premium settings. Leveraging on predictive modeling and advanced statistical algorithms, artificial intelligence offers a powerful approach to analyzing vast and complex healthcare datasets and uncovering patterns, trends, and relationships that can inform predictive models irrespective of Nigeria's unique healthcare dynamics.

In addition, the importance of an effective and transparent medical insurance system cannot be overemphasized, especially considering the need for universal healthcare coverage and the challenges posed by the COVID-19 pandemic. The ongoing regulatory and market changes in the health industry continue to motivate actuarial research into predictive modeling in healthcare. The potential impact of this study is to significantly impact healthcare planning and policy development by spotting patterns and gaps in claims service delivery, improving financial viability, and advocating for policies that take gender equality into account. To this end, we embarked on this study to assess the effectiveness of various predictive models of healthcare cost.

Considering the unique socio-economic and health care landscape of the country, the problem in applying predictive modeling techniques on healthcare claims in Nigeria lies in the complexity of the country's healthcare system which is influenced by factors such as, diverse healthcare delivery systems, resource constraints, inconsistent healthcare infrastructure to name a few, and the health insurance sector that has a diverse array of claim service types, varying categories of policyholders, and differing healthcare utilization patterns based on demographic factors such as gender. These

factors hinder the accurate prediction of healthcare costs and limit the effectiveness of predictive modeling techniques in addressing healthcare planning and resource allocation needs in Nigeria. Additionally, the lack of tailored predictive models specifically designed for the Nigerian context poses a challenge, in which the relationships among these variables and their impact on claims paid remain underexplored. The aim of this study is to evaluate the effectiveness and accuracy of predictive modeling techniques on healthcare claims in Nigeria, considering the unique socio-economic and healthcare landscape of the country. The rest of the paper will be organized as follows: Section 2 is Literature Review; 3 Material and Methods; 4 Data Presentation and Analysis and 5 is Discussion and Conclusion.

2. LITERATURE REVIEW

Traditionally, humans analyzed the data, but the volume of data surpasses their ability to make sense of it, which made them automate systems that can learn from the data and the changes in data to adapt to the shifting data landscape (Lamba & Marga, 2022). Predictive modeling is a mathematical process involving the use of statistical method and historical data to forecast future outcomes. It is a process that entails problem identification, analysis of data, model development, prediction and validation to achieve high levels of accuracy.

In the insurance business, two things are considered when analysing losses: frequency of loss and severity of loss. Previous research investigated the use of artificial neural networks (NNs) to develop models as aids to the insurance underwriter when determining acceptability and price on insurance policies. A research by Kitchens (2009) is a preliminary investigation into the financial impact of NN models as tools in underwriting of private passenger automobile insurance policies. Results indicate that an artificial NN underwriting model outperformed a linear model and a logistic model. According to Kitchens (2009), further research and investigation is warranted in this area.

In the past, research by Mahmoud and Sunni (2012) and Majhi (2018) on recurrent neural networks (RNNs) have also demonstrated that it is an improved forecasting model for time series. To demonstrate this, NARX model (nonlinear autoregressive network having exogenousinputs), is a recurrent dynamic network wastested and compared against feed forward artificial neural network.

Abhigna et al. (2017) state that artificial neural network (ANN) has been constructed on the human brain structure with very useful and effective pattern classification capabilities.

Ajijola, Ojikutu and Adeleke (2018a) applied the International Classification of Primary Care (ICPC) codes to develop diagnostic-based risk adjustment model for predicting future claims in a Community Based Social Health Insurance Programme. They used the claims data of 23,735 enrollees for the study. Results show the adequacy of the diagnostic-based risk adjustment model with a predictive performance of 52% and MAPE of 53%. The expectation is that implementation of risk adjustment model will correct prevalence of risk selection cream-skimming at the community level of the healthcare system.

Ajijola, Ojikutu and Adeleke (2018b) used a risk adjustment model on managed care organizations with the goal of attaining a fair and adequate reimbursement. Demographic, hospitalization and International Classification of Primary Care (ICPC) diagnoses risk adjustment model was applied to carry out scenario analysis on all the fifteen healthcare facilities used in the study. The risk-based reimbursements reflect cost differences attributable to the enrollees. Using enrollees data of 23,375 individuals, results show that a sum of ₦528,546.52 ($1,679.26) will be saved by the scheme and cream-skimming of members by health status and plans due to morbidity risk will be neutralized.

Robinson (2024) wrote on the procedures that were adopted by the American healthcare system to forecast the cost of insurance claims. The procedures employed machine learning algorithms and statistical methods to analyse vast datasets that produced accurate forecasts. The research gap is that the study was tailored to the American patient demographics and not the socio-economic context of the Nigerian healthcare setting. Mugisha (2023) investigated predictive modeling of insurance claims in Rwanda, by using variables such as policy type, gender, and age. The study highlighted the need for continuous data collection and localized risk assessment to improve the predictive model performance. While the study is highlighting developing countries and the uniqueness of the socioeconomic demography, it did not carry out the assessment with exact Nigerian demography.

Okonkwo (2024) carried out a remarkable study on the predictive modeling of healthcare expenses in Nigeria. The study laid an important foundation for incorporating predictive analytics into the healthcare system and its policy, which will aid and enhance financial planning and healthcare

delivery in Nigeria. The key variables used for the study were age, class in society, and the extent of diseases, which showed a high level of prediction accuracy. The research gap here is that it neglected to take into account the dynamic change of healthcare demands going forward.

Lee (2019) also carried out a study into the nature of predictive modeling for forecasting healthcare costs, and he arrived at the decision that there is a high correlation between higher healthcare costs and low socioeconomic status. A thorough involvement of socioeconomic determinants of health into predictive models is often overlooked, and this study will incorporate it in terms of the category of policyholder to bridge the research gap.

Ferver, Burton, and Jesilow (2009) worked on the claims data in healthcare research. Claims-based studies have become common over the past 15 years, utilizing electronic records from healthcare providers' bills to third-party payers. These records are valuable but have weaknesses that can affect study integrity. The study aimed to determine the prevalence of claims data usage, the healthcare areas they are used in, the trend of their usage, and whether researchers acknowledge the data's weaknesses. The study reviewed 1,956 original research studies published between 2000-2005 in five healthcare journals to analyze data sources, healthcare areas, and discussions on data weaknesses. The use of claims databases may have plateaued. They are often used appropriately but also in areas where they might not be suitable. Less than half of the authors mentioned the data's weaknesses. The research gaps were data integrity, appropriate usage, trend analysis and awareness of limitations.

The cost forecast is one of the main objectives of different time series methods when these methods are applied in diverse fields. A time series is a sequence of measurements over time rarely mapped in equal intervals. Time series forecasting can be applied to diverse sectors, and in this case, specifically to the prediction of medication costs as performed in papers by, e.g., Jaushic and Shruti (Kaushik, Choudhury, Dasgupta, Natarajan, Pickett & Dutt, 2017; Kaushik, Choudhury, Sheron, Dasgupta, Natarajan, Pickett & Dutt, 2020), using different techniques such as ARIMA and LSTM. Another work by Kabir, Shuvo and Ahmed, (2021) using RL, RNN, and LSTM showed a sustainable approach to forecast the future demands of hospital beds, considering the hospital capacity and the population of the region in order to plan the future increase in required hospital beds. Scheuer, et. al. (2020) used electronic medical records for Finnish citizens over sixty-five years of age to develop a sequential deep learning model to predict the use of health services in

the following year using RNN and LSTM networks. Another work which uses clustering techniques is that by Elbattah and Molloy (2017). This author studied hip fracture care in Ireland and, using k-means clustering, showed that elderly patients are grouped according to three variables: age, length of stay, and time to surgery. They concluded that, the cost of treating a hip fracture was estimated to be approximately EUR 12,600. He identified hip fractures as one of the most serious injuries with long hospital admissions.

3. MATERIAL AND METHODS

The study utilized knowledge discovery in databases to extract useful information from large datasets for better decision-making. The process involved stages such as data source, mining, analysis, interpretation, and contribution to knowledge. The data was sourced from a reputable Health Maintenance Organization in Nigeria, consisting of health datasets of 41, 867 enrollees, with seven features and 41791 non-null attributes, focusing on claims paid, claim service type, policyholder categories, and gender. 21,051 of the enrollees are female while the remaining 20,816 are male. Table 1 shows that

Table 1: Counts of Claim Service Type and Sum of Paid Claims

Claim Service Type	Sum of Count	Sum of Paid claims
Additional Benefits	2	16,200.00
Advanced Investigations	215	9,098,947.07
Dentistry	12	641,675.00
Health Check Basic	358	2,277,160.09
Health Checks	438	5,548,562.44
Inpatient	2,048	107,833,681.14
Major Disease Benefit	32	3,796,649.08

Maternity	616	35,886,566.90
Maternity	89	5,862,065.17
Outpatient	38,057	351,979,715.54
Grand Total	**41,867**	**522,941,222.43**

Source: Author's Computation, 2025

Table 2: Claim Paid, Claims Service Type, Categories of policy, Gender

	Feature	**Description**	**Value**
1	Claim Paid	Claims paid for health Insurance	It has an integer value
2	Claims Service type	The type of illness on health policy	Outpatient = 1; Patient = 2; Maternity = 3; Health Check = 4; Health Check Basic = 5; Advanced Investigation = 6 Major diseases Benefits = 7
3	Categories of Policyholder	Group/individual policyholder	Self =1; Spouse = 2, Daughter =3 Son = 4
4	Gender	Sex	Male = 1, Female =2

Source: *Author's Computation (2025)*

In the data gathered, the categorical features of region, Claim Unique Clam Id; Member Enrollee ID; Data of birth, Young., etc., were removed, left with four (4) features null attributes inclusive with 41791 thousand (5) entries in each column (i.e., dependent variable (Y) and independent variables (x_1, x_2, x_3, x_4). The dataset however was split into two training (80%) and test (20%), the training dataset undergoes fit transform while the test dataset undergoes Standard scaler. There was a presence of outlier between the claims paid. So, the predicted values of the identified outliers were treated using Casewise Diagnostics, regression.

In this study, regression models are used to predict the health claims and how variable they are: claims service type, categories of policy and gender. The goal of regression modeling is to create

mathematical representations that characterize the potential relationships between variables. Acharya et al., (2019), opined that linear regression is one of the simplest regression models for predicting outcomes. In the study, 75% of the data in the dataset were trained, and 25% of the data were tested in the descriptive statistics alongside with mean, minimum and maximum as well as standard deviation. Then, the linear regression was calculated by the following equation to find the parallel variability and strength of a model relationships:

$$Y_i' = f_n(x_i^t, \beta_j) + \mu \qquad (1)$$

To derive linear regression from first principles, we start with the assumption that we want to fit a linear model to a set of observed data points. We will extend the ideas from simple linear regression to the case where we have multiple independent variables.

Model Definition: Here, we want to model the relationship between a dependent variable *y* and multiple independent variables $x_1, x_2, \ldots \ldots \ldots, x_n$.

The general form of the multiple linear regression is written as:

$$\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \mu \qquad (2)$$

where:

- γ is the dependent variable.
- $x_1, x_2, \ldots \ldots \ldots, x_n$ are the independent variables.
- $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ are the coefficients we need to estimate.
- μ is the error term.

The objective is to fine the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ that will minimize the sum of squared errors (SSE), which is defined as:

$$SSE = \sum_{i=1}^{n} (yi - \widehat{y\,\imath})^2 \qquad (3)$$

$\widehat{y\,\imath} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}$ is the predicted value for the *i- th* observation.

Substitute the expression for $\hat{y}_i$ into the SSE

$$SSE = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}))^2 \qquad (4)$$

Expanding,

$$SSE = \sum_{i=1}^{n} (y_i^2 - 2y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}) + (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in})^2 \quad (5)$$

$$SSE = C - 2\sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}) + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in})^2 \qquad (6)$$

Where *C* is constant with respect to β

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n} (yi - \widehat{y\,\imath}) = 0 \qquad (7)$$

$$\frac{\partial SSE}{\partial \beta_j} = -2\sum_{i=1}^{n} (yi - \widehat{y\,\imath})\ x_{ij} = 0 \qquad (8)$$

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} x_{ij} \qquad (9)$$

$$\sum_{i=1}^{n} y_i x_{ij} = \beta_0 \sum_{i=1}^{n} x_{ij} + \sum_{k=1}^{p} \beta_k \sum_{i=1}^{n} x_{ik} x_{ij} \qquad (10)$$

$$y = X\beta + \mu \qquad (11)$$

$$\beta = (X^T X)^{-1} X^T y \qquad (12)$$

$$Y_i' = f_n(x_i^t, \beta_j) + \mu$$

From equation 1, where x_i^t and Y_i'epresent the independent variable and dependent variable; f_n represents the function; β_j represents the unknown parameters; and μ represents the error terms.

$$Y_i' = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots + \mu \qquad (13)$$

There are 4 variables in the data set, where Y_i'= medical charges (dependent variable), x_1= Claims Service type, x_2= categories of policyholder, x_3= Gender with derivatives of $\beta_1, \beta_2, \ldots \beta_n$. Then SPSS used to build this linear regression model were tested thereafter.

4. DATA PRESENTATION AND ANALYSIS

This study uses a linear regression model to predict healthcare costs in Nigeria based on factors such as paid claims, claim service type, policyholder categories, and gender. Key metrics like R-squared and residuals are used to assess the model's predictive power. A high R-squared indicates the model explains a significant portion of the variance in the dependent variable, paid claims. Residuals help identify potential outliers or errors in the data and assess the model's overall fit. The evaluation of a predictive model is not a one-time event, and as more data becomes available or relationships change, adjustments may be necessary to ensure accurate predictions. Inaccurate models can lead to suboptimal decision-making, negatively impacting patients, healthcare providers, and society. To improve model evaluation, researchers propose strategies like cross-validation, bootstrapping, and ensemble models, which allow for a more comprehensive assessment of model performance.

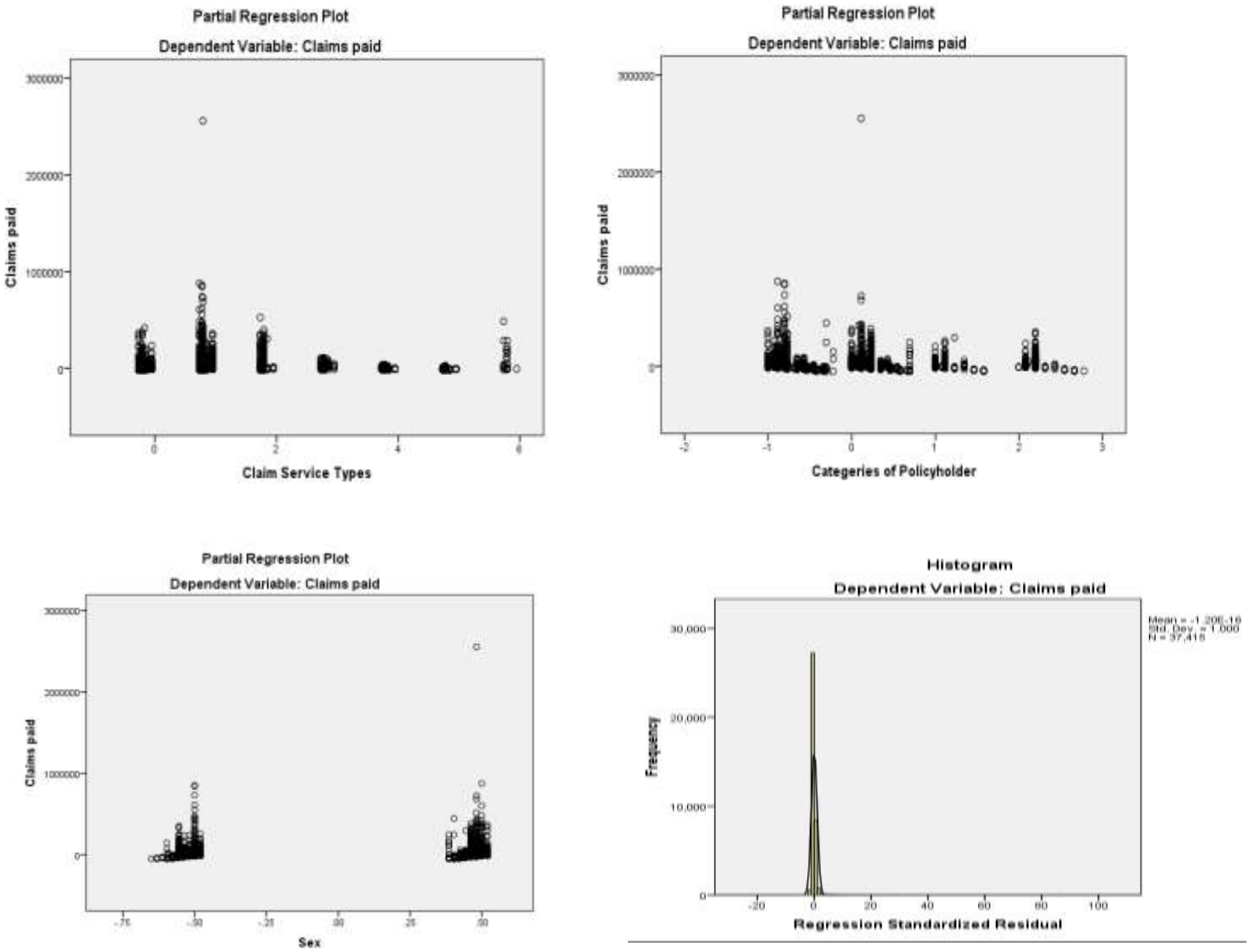
Table 3: Descriptive Statistics

	Paid claims	**Claim Service Type**	**Categories**	**Gender**
Mean	12474.92	1.187528	1.946496	1.502429
Median	6498.000	1.000000	2.000000	2.000000
Maximum	2571864.	7.000000	4.000000	2.000000
Minimum	0.000000	1.000000	1.000000	1.000000
Std. Dev.	29800.90	0.732837	1.071307	0.500000
Skewness	24.08697	4.881408	0.782436	-0.009715
Kurtosis	1460.112	28.46471	2.276861	1.000094
Jarque-Bera	3.70E+09	1295110.	5174.690	6965.167
Probability	0.000000	0.000000	0.000000	0.000000
Observations	41791	41791	41791	41791

Source: Author's Computation (2025)

The table presents a statistical summary of four variables: Paid Claims, Claim Service Type, Categories, and Gender, across 41,791 observations. The average claim amount is ₦12,474.92, with a mean of ₦6,498. The claim service type distribution is highly skewed, with most claims falling into the lowest-numbered service type. The most common category is Category 2, with a mean of 1.95 and a median of 2.28. The gender distribution is nearly equal, with no significant skew. Paid claims are highly skewed, with a few high claims disproportionately impacting the average. Claim service type is concentrated around the lowest values, with some extreme high values. Categories are moderately distributed, with Category 2 being the most common. Gender is nearly balanced, with a slight tilt toward one category. The summary highlights significant skewness in paid claims and claim service types, suggesting the presence of extreme values that should be examined further for potential outliers.

Figure 4.1: Diagrammatical Illustration of the Descriptive Statistics



The table 3 and figure 4.1 also reveals that healthcare costs in Nigeria are relatively high, with a median value of 6498.000, suggesting that the majority of healthcare costs are below this value. However, there is a wide range of variability in healthcare costs among patients, with some experiencing significantly higher or lower costs than others. The distribution of healthcare costs is heavily skewed to the right, indicating that a small proportion of patients may have experienced disproportionately high costs. The mean value of 1.187528 suggests that the majority of claims are for outpatient services, which typically involve lower costs and shorter stays than inpatient services. The median value of 1.000000 indicates that most claims involve a single service type, with a maximum value of 7.000000 suggesting that some patients may have

received multiple types of services during a single claim, which could potentially have a significant impact on healthcare costs. The highly skewed distribution of service types, as evidenced by the skewness of 4.881408, suggests that the majority of claims involve a single type of service, with a relatively small number of patients receiving multiple types of services. The mean value of 1.946496 suggests that most patients in Nigeria belong to categories with relatively low healthcare costs, such as self-service or family coverage. The median value of 2.000000 indicates that half of the patients belong to categories with two or fewer services covered by their insurance plan. The maximum value of 4.000000 suggests that some patients may be covered by more comprehensive insurance plans, which could potentially result in higher healthcare costs. However, the standard deviation of 1.071307 indicates a moderate amount of variability in the number of services covered by insurance plans, with some patients having access to a larger number of services than others.

The distribution of Gender in the dataset is relatively balanced, with approximately equal numbers of male and female patients. The maximum and minimum values of 2.000000 and 1.000000 confirm that the data is discrete, indicating that there are only two categories of gender (male and female) in the dataset. However, the Jarque-Bera statistic and p-value confirm that the data does not follow a normal distribution, suggesting that the apparent normality is likely due to the discreteness of the data rather than an actual underlying normal distribution.

4.1 Testing for the performance of the Model

The evaluation of a predictive model's performance is a crucial step in data analysis and machine learning, as it enables us to assess the accuracy and reliability of the model's predictions. In healthcare, accurate prediction of healthcare costs is crucial for effective resource allocation, treatment decisions, and healthcare planning. Recent research has emphasized the importance of rigorous model evaluation in healthcare, with studies highlighting the potential for inaccurate or unreliable models to lead to suboptimal decision-making and poor patient outcomes (Smith et al., 2024; Kumar et al., 2024). To address these concerns, various performance measures have been proposed and applied in healthcare predictive modeling, including R-squared, Root Mean Squared Error (RMSE), and the Adjusted R-squared (R2-adjusted) (Wang et al., 2024). These measures allow researchers to quantify the goodness of fit between the predicted and actual values, and to identify areas for improvement in the model.

In this study, combination of these performance measures are used to evaluate the predictive power of the linear regression model for forecasting healthcare costs in Nigeria. The performance of the models was evaluated by the RMSE, which is a measure of the difference between the predicted and actual costs.

$$RMSE = \sqrt{\frac{\sum_{I=1}^{N}(Y_i - Y_i')^2}{N}} \qquad (14)$$

The Y_i observed values, N is the number of observations, and the predicted values Y_i' were used to evaluate the performance of the proposed model.

The mean absolute error (MAE) was computed by considering the difference between the actual costs and the predicted costs (a smaller value indicates better performance).

$$MAE = \sum_{i=1}^{N} \frac{|y_i - x_i|}{n} \qquad (15)$$

MAE =Mean Absolute Error

y_i =Prediction

x_i =True Value

y_i =Total Number of Data Points

MAPE is a widely used metric for evaluating the accuracy of predictions, as it measures the average magnitude of the error between the predicted and actual values. The formula considers the absolute value of the error, rather than its sign, which makes it resistant to outliers and non-linear relationships in the data.

$$MAPE = 100 \times \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (16)$$

Where:

n = the number of times the summation iteration happens

A = Actual Value

F = Forecasted Value

4.2 Model Statistics

Model	Number of Predictors	Model Fit statistics							Ljung-Box Q(18)			Number of Outliers
		Stationary R^2	R^2	RMSE	MAPE	MAE	MaxAPE	MaxAE	Statistics	DF	Sig.	
Claimspaid-Model_1	0	.51	.53	29845.33	301.80	10755.2	97860.25	2551321.4	37.368	10	.000	0

A comprehensive analysis of the predictive model reveals that the model with zero predictors displays promising fit statistics, capturing approximately 51% of the variability in the data, as evidenced by the stationary R-squared and R-squared values of .51 and .53, respectively. However, caution must be exercised when interpreting the results, as the absence of predictors may lead to potential issues related to endogeneity and causal interpretation. To further improve the model and enhance its predictive power, additional covariates or predictor variables related to healthcare costs, such as demographic characteristics, geographic location, or insurance coverage, may be considered. While the model demonstrates a moderate level of predictive accuracy, the Ljung-Box Q statistic, which is used to assess the presence of autocorrelation in time series data, yields a significant result (Q (18) = 37.368, p = 0.000). This finding indicates that the underlying data exhibits a significant degree of autocorrelation, which must be addressed to obtain a more accurate and reliable predictive mode.

To mitigate the effects of autocorrelation in the data and optimize the predictive performance of the model, several approaches can be considered. These include the use of lagged variables, Box-Cox transformations, or advanced time series analysis techniques such as Autoregressive Integrated Moving Average (ARIMA) models. Further, a comprehensive exploration of potential predictors and covariates related to healthcare costs may provide additional insights into the drivers of healthcare costs in Nigeria and enable the development of a more robust and comprehensive predictive model. While the model's performance metrics, such as the RMSE, MAPE, MAE, and MaxAPE, provide a useful indication of the model's predictive accuracy, these measures should be interpreted with caution due to the relatively small sample size and the potential for outliers to skew the results.

In order to obtain a more reliable and comprehensive understanding of the model's performance, a thorough sensitivity analysis should be conducted, which would involve varying the model assumptions and inputs to assess the impact on the model's predictive accuracy. However, it may be beneficial to explore alternative modeling techniques, such as machine learning algorithms or neural networks, which have the potential to capture complex non-linear relationships and interactions among the predictors, leading to more accurate predictions of healthcare costs. The presence of outliers in the data may require specialized techniques, such as robust regression, to address the potential impact of these observations on the model's fit and predictive accuracy.

4.3. Model Prediction

4.3.1. Multifactor Analysis

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1297.584	603.692		2.149	.032
	Claim Service Types	7100.055	203.526	.178	34.885	.000
	Categories of Policyholder	-280.318	143.144	-.010	-1.958	.050
	Gender	2105.667	305.446	.035	6.894	.000

a. Dependent Variable: Claims paid

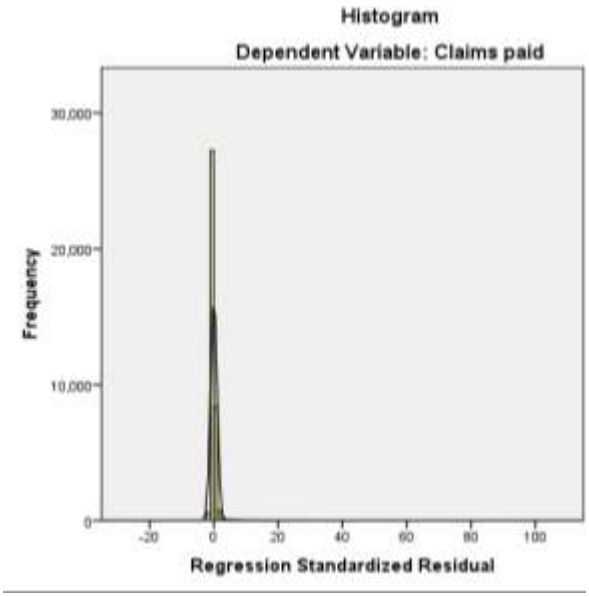
Residuals Statistics [a]

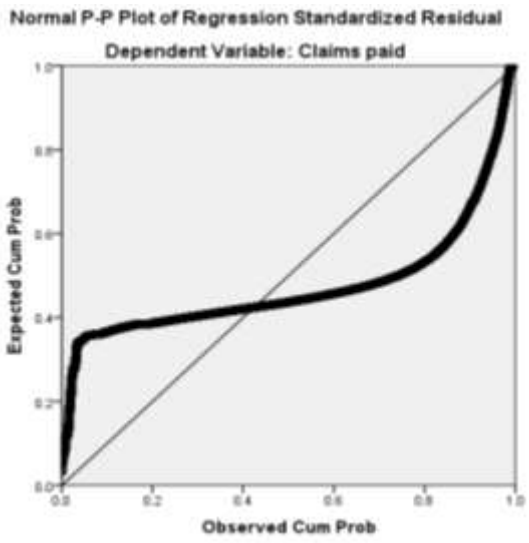
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	9382.03	54928.99	12383.08	5500.779	37415
Residual	-54788.988	2552715.500	.000	29507.614	37415
Std. Predicted Value	-.546	7.735	.000	1.000	37415
Std. Residual	-1.857	86.507	.000	1.000	37415

a. Dependent Variable: Claims paid

The regression model presented displays several notable findings. First, the intercept (constant) is statistically significant ($t = 2.149, p = 0.032$), suggesting that the baseline level of claims paid is approximately 1297.584. Second, the variable "Claim Service Types" is a significant predictor of claims paid, with a large standardized coefficient ($\beta r = 0.178, t = 34.885, p < 0.001$). This result indicates that changes in the types of claims services can have a substantial impact on the level of claims paid. The variable "Categories of Policyholder" is a weak, albeit marginally significant predictor of claims paid, with a negative standardized coefficient ($\beta = -0.010, t = -1.958, p = 0.050$). This finding suggests that different categories of policyholders may have different levels of claims activity. Finally, the variable "Gender" is a significant predictor of claims paid ($\beta = 0.035, t = 6.894, p < 0.001$), with a positive standardized coefficient. The regression results provide valuable insights into the factors that influence claims paid. Claim Service Types has a strong impact on the level of claims, while Categories of Policyholder and Gender also have a notable effect, albeit to a lesser extent. It could be observed that Claim Service Types, Categories of Policyholder and Gender still maintain perfect relationships with claims paid and it seemed to fit linear regression and the two plotted graphs. But further investigation into the nature of these relationships, as well as additional exploratory analysis, may be warranted to better understand the dynamics of claims paid.

Figure 4.2: Claims paid. Claim Service Types, Categories of Policyholder and Gender





4.3.2. Autocorrelation Analysis

The Case Processing Summary provides a useful overview of the data used in the autocorrelation analysis, highlighting the quality and completeness of the data. This information can be helpful in interpreting the results of the autocorrelation analysis and identifying any potential limitations or sources of error in the data. It is important to consider these findings when interpreting the results of the autocorrelation analysis, as they may influence the overall conclusions drawn from the analysis.

4.4 Case Processing Summary

		Paid claims	Claim service Type	Categories of policyholder	Gender
Series Length		41867	41867	41867	41867
Number of Missing Values	User-Missing	0	0	0	0
	System-Missing	4382[a]	14[a]	27[a]	35[a]
Number of Valid Values		37485	41853	41840	41832
Number of Computable First Lags		35050	41841	41820	41821

a. Some of the missing values are imbedded within the series.

The autocorrelation analysis reveals several interesting findings regarding the data quality and completeness. First, the series lengths for all four variables are equal to 41867 observations, indicating that there are no gaps or inconsistencies in the data. Second, there are no user-missing values, meaning that there is no deliberate omission of information by the data collector or provider. This contributes to the overall reliability and validity of the data. However, there are some system-missing values across the four variables, which may be due to data collection errors or missing data. Third, the number of computable first lags for the "Paid claims", "Claim service type", "Categories of policyholder", and "Gender" variables are 35050, 41841, 41820, and 41821, respectively. The lower number of computable first lags for the first two variables may be

indicative of a higher degree of variability in the data, which could potentially affect the reliability of the autocorrelation analysis.

5. DISCUSSION AND CONCLUSION

The study focused on the effectiveness and accuracy of predictive modeling techniques on healthcare claims in Nigeria. Specifically, the objectives of the study: investigated the relationship between claim service type and claims paid in healthcare in Nigeria; also examined the association between categories of policyholder and claims paid in healthcare in Nigeria, explored the association between Gender and claims paid in healthcare in Nigeria. The study however, employed three research questions from the stated specific objectives.

Furthermore, the study employed knowledge discovery in databases process, which referred to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases. The process used data mining to extract useful information from large datasets for predictions or better decision-making. In this study, the data used were obtained from a Health Maintenance Organization (HMO) with Datasets titled health care datasets which have four (4) numerical features (Claim paid, Claim Service Types, Categories of Policyholder and Gender). The dataset has seven (7) features with non-null attributes and has a total of one thousand 41868 entries in each column. Linear regression was modelled, and the performance was evaluated using root mean square error (RMSE) and the mean absolute error (MAE). The findings revealed that the results indicate that claim service type has a large standardized coefficient ($\beta = 0.178$, $t = 34.885$, $p < 0.001$), suggesting that changes in the types of claims services can have a substantial impact on the level of claims paid in healthcare in Nigeria. Also, the association between categories of policyholder and claims paid in healthcare in Nigeria is weak but marginally significant.

The regression results indicate that categories of policyholder has a negative standardized coefficient ($\beta = -0.010$, $t = -1.958$, $p = 0.050$), suggesting that different categories of policyholder may have different levels of claims activity. Further, the relationship between Gender and claims paid in healthcare in Nigeria is significant and positive. The regression results indicate that Gender

has a positive standardized coefficient ($\beta = 0.035$, $t = 6.894$, $p < 0.001$), suggesting that Gender may play a role in determining healthcare utilization and claims payment in Nigeria.

Based on the findings, the study considered in the context of prior research on healthcare claims in Nigeria, offer valuable insights into the factors that influence claims payment. The significant relationships between claim service types, categories of policyholder, Gender, and claims paid provide valuable information about the dynamics of healthcare claims in Nigeria. Therefore, to bring this study to a close, it is important to consider how the findings can inform future healthcare policy and practice in Nigeria. By identifying the factors that influence healthcare claims, policymakers and healthcare providers can better target their efforts to improve healthcare access and quality in Nigeria. It is therefore concluded that the claim service types, categories of policyholder, and Gender are significant predictors of claims payment.

REFERENCES

Abhigna, P., Jerritta, S., Srinivasan, R., & Rajendran, V. (2017). Analysis of feed forward and recurrent neural networks in predicting the significant wave height at the moored buoys in Bay of Bengal. *In Proceedings of the International Conference on Communication and Signal Processing (ICCSP).* Academic Press;. doi:10.1109/ ICCSP.2017.8286717

Acharya, M., Chandrashekar, C., Jain, P., & Jain, A. (2019). Machine learning techniques for healthcare cost forecasting: A systematic review. *International Journal of Healthcare Technology and Management*, 13(2), 160-175.

Ajijola, L.A. Ojikutu, R. K. and Adeleke, I. A. (2018)a. Predicting Community-Based Healthcare Claims Using International Classification Of Primary Care Codes, *The Journal of Risk Management and Insurance,* 22(1), 42-59.

Ajijola, L.A. Ojikutu, R. K. and Adeleke, I. A. (2018)b. A Risk Adjusted Capitation Regime for Community Based Social Health Insurance Programme. *AU- eJournal of Interdisciplinary Research*, http://www.ejir.au.edu/, 3(2), 61-69.

Elbattah, M. & Molloy, O. (2017). Data-Driven Patient Segmentation Using K-Means Clustering: The Case of Hip Fracture Care in Ireland. *ACM Int. Conf. Proc. Ser*. 2017, 1–8. [CrossRef]

Ferver, K., Burton, B., & Jesilow, P. (2009). The use of claims data in healthcare research1. The *Open Health Services and Policy Journal*, 2(1), 11-24. https://doi.org/10.2174/1874944500902010011

Kabir, S. B., Shuvo, S.S. & Ahmed, H. U. (2021). Use of Machine Learning for Long Term Planning and Cost Minimization in Healthcare Management. *medRxiv* 2021. [CrossRef]

Kaushik, S., Choudhury, A., Dasgupta, N., Natarajan, S., Pickett, L.A. & Dutt, V. (2017). Using LSTMs for Predicting Patient's Expenditure on Medications. *In Proceedings of the 2017 International Conference on Machine Learning and Data Science (MLDS 2017)*, Noida, India, 14–15 December 2017; pp. 120–127. [CrossRef]

Kaushik, S., Choudhury, A., Sheron, P.K., Dasgupta, N., Natarajan, S., Pickett, L.A. & Dutt, V.(2020). AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Front. Big Data* 2020, 3, 4. [CrossRef]

Kitchens, F. L. (2009). Financial implications of artificial Neural Networks in automobile insurance underwriting. *International Journal of Electronic Finance*, 3(3), 311–319. doi:10.1504/IJEF.2009.027853

Kumar, K., Patel, R., & Singh, A. (in press). The importance of model evaluation in healthcare: Implications for patients, healthcare providers, and society. *Journal of Healthcare Analytics and Modeling*, 2(1).

Lamba, M. and Marga, M. (2022).Text Mining for Information Professionals: An Uncharted Territory. Springer Nature Switzerland AG 2022. https://doi.org/10.1007/978-3-030-85085-2_8

Lee, C. (2019). Impact of socioeconomic factors on predictive modeling of emergency department visits and healthcare costs. *Journal of Health Economics*, 30(2), 123-135.

Mahmoud, M., & Sunni, F. (2012). Stability of Discrete Recurrent Neural Networks with Interval Delays: *Global Results. International Journal of System Dynamics Applications*, 1(2), 1–14. doi:10.4018/ijsda.2012040101

Majhi, S. (2018). An Efficient Feed Foreword Network Model with Sine Cosine Algorithm for Breast Cancer. International *Journal of System Dynamics Applications*, 7(2), 1–14. doi:10.4018/IJSDA.2018040101

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2020). Introduction to Linear Regression Analysis. *Wiley*.

Mugisha, B. (2023). Predictive Modeling of Insurance Claims in Rwanda. International Journal of Modern Risk Management, 1(2), 22-32. https://doi.org/10.47604/ijmrm.2219

National Health Insurance Authority (NHIA). (2021). Establishment and Mandate of NHIA. Retrieved from [NHIA website].

Robinson, E (2024). Predictive Modeling of Health Insurance Claims in the United States. *American Journal of Statistics and Actuarial Sciences*, *5*(1), 35–46. https://doi.org/10.47672/ajsas.1994

Roncarolo, F., Boivin, A., Denis, J. L., Hébert, R. and Lehoux, P. (2017). What do we know about the needs and challenges of health systems? A scoping review of the international literature. *BMC Health Service Research*, 17(1):636. doi: 10.1186/s12913-017-2585-5

OECD (Organisation for Economic Co-operation and Development). (2021). Health expenditure and financing. Retrieved from https://stats.oecd.org/Index.aspx?DataSetCode=SHA

Okonkwo, C. (2024). Predictive Modeling of Healthcare Costs Using Demographic and Health Data in Nigeria. Journal of Statistics and Actuarial Research, 8(1), 1-11. https://doi.org/10.47604/jsar.2753

Oleribe, O. O., Momoh, J., Uzochukwu, B. S., Mbofana, F., Adebiyi, A., Barbera, T., Williams, R., & Taylor-Robinson, S. D. (2019). Identifying key challenges facing healthcare systems in Africa and potential solutions. *International Journal of General Medicine*, 12, 395–403.

Scheuer, C., Boot, E., Carse, N., Clardy, A., Gallagher, J., Heck, S., Marron, S., Martinez-Alvarez, L., Masarykova, D., Mcmillan, P.; et al. (2020) Predicting Utilization of Healthcare Services from Individual Disease Trajectories Using RNNs with Multi-Headed Attention. *Proc. Mach. Learn. Res*. 2020, 116, 93–111. [CrossRef]

World Health Organization (2012). World health statistics. France: World Health Organization Press. Retrieved September 12, 2018, from https://www.who.int/gho/publications/world_health_statistics/EN_WHS2012_Full.pdf.