Using Principal Component Analysis and Logistics Regression to Model Major Types of Cancer among Youth: A Review

Mansur Adamu Abubakar 1; M. U. Adehi 2; and M. A. Auwal 2

1Nigeria Incentive-Based Risk Sharing System for Agricultural Lending, 2Department of Statistics, Nasarawa State University, Keffi Nigeria.

Corresponding author: m.abubakar@nirsal.com

DOI: https://doi.org/10.62154/ee3ykg53

Abstract
Cancer poses a significant health challenge among young Nigerians due to high incidence rates and limited research on risk factors and outcomes. This study examines the utility of logistic regression and principal component analysis (PCA) in understanding cancer patterns, identifying risk factors, and predicting outcomes among Nigerian youth. Logistic regression assesses various factors such as genetics, environment, and lifestyle to predict cancer occurrence, enabling early detection and tailored interventions. PCA aids in uncovering complex data patterns by reducing dimensionality while preserving data variability. Nigeria experiencing over 100,000 new cancer cases annually and a rising burden, effective statistical modeling is vital for guiding prevention and intervention strategies. The study compares different cancer types, develops risk assessment models, and forecasts cancer probabilities using logistic regression and PCA. These methodologies hold promise in enhancing cancer comprehension, diagnosis, and management among youth. Patient cancer data from the National Hospital in Abuja and the Aminu Kano Teaching Hospital in Kano were utilized, selected through convenient sampling for accessibility and convenience. Direct observation served as the primary data collection method. This research underscores the importance of employing statistical modeling techniques, particularly logistic regression and PCA, to address the multifaceted challenges of cancer among youth. Leveraging analytical frameworks can advance early detection initiatives, tailored prevention strategies, and timely interventions, aligned with the region’s unique socio-economic and cultural dynamics. Recommendations include prioritizing model validation, fostering collaborative efforts, increasing investment in cancer research infrastructure, and leveraging statistical modeling techniques to enhance cancer epidemiology understanding and healthcare outcomes among youth.

Keywords: Principal Component Analysis, World Health Organization, Low- and Middle-Income Countries, Logistic Regression Model, Logistic Regression.

Introduction
In cancer research among youth in Nigeria, logistic regression, and principal component analysis (PCA) serve as valuable statistical tools for analyzing and understanding complex data patterns associated with cancer incidence, risk factors, and outcomes. Logistic
regression allows researchers to examine the relationship between various risk factors (such as genetic predisposition, environmental factors, and lifestyle choices) and the likelihood of cancer occurrence among young individuals. By modeling the probability of cancer diagnosis as a binary outcome, logistic regression helps identify significant predictors and quantify their impact on cancer risk, thereby facilitating early detection, prevention, and intervention strategies tailored to the Nigerian youth population. On the other hand, PCA enables researchers to explore the multidimensional nature of cancer-related data by reducing the dimensionality of the dataset while preserving most of its variability.

Cancer is one of the top causes of death worldwide, accounting for an estimated 8.2 million deaths in 2018 (World Health Organization [WHO], 2018). In 2012, there were around 32.6 million reported cancer cases worldwide, with 14.1 million new cases described in the literature (Salako et al., 2018). Following this trend, an estimated 16 million additional cases per year are projected in 2020, with developing nations accounting for more than 70% of these cases (Sowunmi et al., 2018). In 2018, a global update found that there were around 18.1 million new cases and approximately 9.6 million deaths (WHO, 2018).

According to current data in Nigeria, over 200,000 new cases of cancer are diagnosed each year, with a high case-fatality ratio. This growing cancer burden in Sub-Saharan Africa (SSA), particularly Nigeria, has created health issues for residents and the government (Jedy-Agba et al., 2012). Because of the morbidity impacts of other illnesses and ailments, nearly 70% of cancer cases have at least one comorbidity (Bellizzi & Rowland, 2007). However, these comorbidities have made cancer therapy extremely challenging, negatively impacting patients' survival and quality of life. The Abuja cancer registry showed an aged, standardized incidence rate of 58.3/100,000 men and 138.6/100,000 women (Jedy-Agba et al., 2012; Sowunmi et al., 2018).

Statement of the Research Problem
According to the World Health Organization (WHO), cancer is a leading cause of morbidity and mortality worldwide, with approximately 70% of cancer deaths occurring in low- and middle-income countries (LMICs) like Nigeria (WHO, 2018). Cancer is a leading cause of death in children and adolescents worldwide (WHO, 2018). In Nigeria, cancer is the second leading cause of death in children under the age of 15 years. The most common types of cancer in youth in Nigeria include leukemia, lymphoma, brain and central nervous system tumors, bone cancer, and soft tissue sarcoma.

Despite the high prevalence of cancer in youth in Nigeria, there is a lack of research on the risk factors for and predictors of cancer in this population. This is due to several factors, including the limited availability of data, the lack of trained researchers, and the paucity of funding for cancer research (Olaniyi et al., 2022). Statistical modelling is a powerful tool that can be used to identify risk factors for cancer, predict the likelihood of cancer development, and assess the effectiveness of cancer treatments. Logistic regression is a
statistical model that can be used to predict binary outcomes, such as the presence or absence of cancer. PCA is a statistical technique that can be used to reduce the dimensionality of data and identify underlying patterns. This review is essential given the increased cancer incidence among young in Nigeria. It seeks to provide a greater knowledge of the variables causing youth cancer and make it easier to establish tailored prevention interventions by incorporating cutting-edge statistical tools. Therefore, the problem statement for this research is: how can machine learning and statistical modelling be used to accurately predict and diagnose different types of cancer among youth in Abuja and Kano, considering the unique socio-economic and cultural factors of the country, and how can these models be validated and tested in real-world settings.

Objectives of the Review of Statistical Modelling of Major types of cancer
1. To compare and classify major types of cancer among youth using Logistic Regression and Principal Component Analysis (PCA)
2. Develop risk assessment and early diagnosis model of major types of cancer among youth using Principal Component Analysis (PCA)
3. Develop a classifier model using logistics regression and Principal Component Analysis (PCA) for each of the specified conditions.
4. Utilise logistics regression models to predict the probability of youth developing specific types of cancer-based on the reduced set of influential variables.

Characteristics of Logistic Regression
i. Binary Outcome Prediction: Logistic regression is characterized by its ability to model the relationship between one or more independent variables and a binary outcome. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of occurrence of one of two possible outcomes, typically represented as 0 or 1.
ii. Probability Estimation: Logistic regression employs the logistic function (sigmoid function) to estimate the probability of the occurrence of the event of interest. This function maps any real-valued input into a range between 0 and 1, making it suitable for modeling probabilities. The predicted probabilities are then used to make binary classifications based on a chosen threshold.
iii. Assumption of Linearity in Logit: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. The logit transformation is applied to convert the odds ratio into a linear relationship, facilitating the interpretation of the coefficients associated with each independent variable.
iv. Maximum Likelihood Estimation: Logistic regression utilizes maximum likelihood estimation to estimate the parameters of the model. The algorithm iteratively adjusts the coefficients to maximize the likelihood of observing the actual outcomes given the predictors, ultimately finding the set of coefficients that best fit the data.

v. Goodness-of-Fit Measures: Logistic regression employs various measures, such as the likelihood ratio test, Wald test, and Hosmer-Lemeshow test, to assess the goodness of fit of the model. These measures evaluate how well the model fits the observed data and provide insights into its predictive performance and reliability.

vi. Applications: Logistic regression finds applications in various fields, including medicine, economics, marketing, and social sciences. It is commonly used for binary classification tasks such as predicting the likelihood of disease occurrence, customer churn, loan default, and more.

These characteristics collectively make logistic regression a versatile and widely used statistical method for binary classification and probability estimation tasks in both research and practical applications.

Literature Review
Review Article Based on the Study
Ayer et al. (2010) sought to examine and compare these two famous models—logistic regression and ANN—in the context of assessing breast cancer risk. This study, which focused on mammographic descriptors and demographic risk variables, aimed to clarify the complex strengths and limits of each strategy, finally providing insights into their value and relevance in clinical decision-making processes.

Ayer et al. (2010) conducted a comparative analysis that provided light on the dynamic landscape of computer-aided models in medical diagnosis, highlighting the critical significance of logistic regression and artificial neural networks in evaluating breast cancer risk. By explaining the distinct characteristics of each strategy and investigating their possible synergies, this study provides useful insights that may inform future research endeavors and clinical applications in oncology and beyond (Ayer et al., 2010).

In an attempt to improve classification accuracy, data-driven methods have been investigated to find consistent differentiating factors between tumor types as well as within-group changes. To clarify the complex statistical discrepancies between tumors, a recent study used gene expression data, with the goal of connecting these differences to phenotypic differences in illness outcomes (Chavez et al., 2021). Using dimensionality reduction techniques like Principal Component Analysis (PCA) and a panel of 6567 genes, researchers were able to identify significant patterns of gene expression linked to different forms of cancer (Chavez et al., 2021).

A thorough examination of the gene expression profiles of different cancer kinds, providing insightful information about the statistical parallels and discrepancies between the
deadliest cancer types (Chavez et al., 2021). The foundation for further research into the biological consequences of gene expression profiles has been laid by the identification of significant genes linked to cancer types by researchers using network-based methodologies and sophisticated data analysis techniques. To further our knowledge of cancer biology and enhance clinical outcomes, our work emphasizes the significance of combining gene expression data with standard statistical methods (Chavez et al., 2021).

Algamal and Lee (2015) state that although cancer categorization is an important use of gene expression analysis, there are substantial obstacles due to the high complexity of genetic data. Adaptive regularized logistic regression methods, specifically AA ELastic, provide a viable approach to gene selection and gene coefficient estimate at the same time. Among the several techniques for accurately classifying tumors in high-dimensional datasets, AA Elastic stands out for its emphasis on grouping effects and consistency in gene selection. This study emphasizes how crucial sophisticated computational techniques are to expanding our knowledge of cancer biology and enhancing our capacity for diagnosis and prognosis (Algamal & Lee, 2015).

A study by Chao, Yu, Cheng, & Kuo (2014) in Taiwan used decision tree methods, logistic regression, and support vector machine (SVM) to create a thorough classification model for the survival rates of breast cancer patients. Utilizing a dataset of 1,340 cases of breast cancer patients from a particular Central Taiwanese hospital, the study aimed to offer physicians a trustworthy framework for evaluating the prognosis of breast cancer patients. All things considered, the study emphasizes how important data mining methods are to enhancing our comprehension of the dynamics of breast cancer survival and giving physicians useful predictive instruments (Chao, Yu, Cheng, & Kuo, 2014).

Wang et al. (2015) investigated applying data mining techniques on several datasets to determine the most effective approach for breast cancer forecasts. They used AdaBoost Tree, naïve Bayes classifier, artificial neural network (ANN), and support vector machine (SVM). After talking about reducing the feature space, Principal Component Analysis (PCA) was used to try and achieve reduction (Wang et al., 2015).

Oyewola et al. (2017) investigated the correlation between breast cancer biopsy forecasts and mammographic diagnosis. The classification methods employed in their investigation were logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), random forest (FR), and support vector machine (SVM) (Oyewola et al., 2017).

Westerdijk (2018) investigated various machine learning algorithms for the prediction of breast cancer cells. She evaluated the models’ performance by examining their accuracies, sensitivity, and specificity. The accuracy scores for LR, random forest, SVM, neural network, and ensemble models were compared (Westerdijk, 2018).

In the context of anticipating linear relationships among variables, PCA doesn’t involve model evaluation. The notion of perceiving data from various standpoints aids in comprehension. Through a transformation involving the relocation of the data’s centroid as
the new origin and subsequent rotation of axes, PCA computes \((PCI,...,PCq)\) based on a set of \(q\) variables \((X_1,...Xq)\) (Jolliffe, 2014). This computation ensures that the new variables, designated as principal components (PCs), encapsulate an equivalent overall variance as the original variables. Principal Component Analysis (PCA) stands as a pivotal statistical method with widespread applicability across various domains, notably including facial recognition (Turk & Pentland, 1991). PCA functions by unveiling the underlying covariance structures among variables within datasets, thereby illuminating the dataset's inherent structure. Fundamentally, PCA endeavors to discern the intricacies embedded within the variance-covariance matrix. Practically, it effectively identifies variables or clusters of variables exhibiting notable interrelationships or high correlations (Jolliffe, 2002).

Principal Component Analysis (PCA) is employed on a singular dataset without distinguishing between predictor and response variables. Specifically in multiple linear regression (MLR), such as Logistic Regression with multiple variables, PCA is solely utilized on the set of \(X\) variables to explore the correlations between pairs of variables (Kuhn & Johnson, 2013).

Methodology

Descriptive research design was used in this study to formulate patient cancer data, which will serve as the focal point of investigation. The hospitals under examination include the National Hospital in Abuja and the Aminu Kano Teaching Hospital in Kano. The case study hospitals were chosen through a convenient sampling method, a non-probability sampling technique. This method involves selecting participants based on their accessibility and convenience rather than using random or systematic sampling. It entails choosing hospitals that are readily available and easily accessible. The data collection process involved direct observation, constituting primary data collection.

Model Interpretation

The logistic regression coefficients can be interpreted as the change in the logit of the dependent variable for a one-unit increase in the corresponding independent variable. For example, if the logistic regression coefficient for a particular independent variable is 0.5, then a one-unit increase in that independent variable would increase the logit of the dependent variable by 0.5.

The common definition of Logistic Function is as follows:

\[
P(x) = \frac{1}{1 + \exp(-x)}
\]  
(1)

Where \(x \in R\) is the variable of the function and \(P(x) \in [0,1]\) one importance of equation (1) is that.

\[
P(-x) = \frac{1}{1 + \exp(x)}
\]  
(2)
The form of equation 3.1 is widely used as the form of logistic regression.

\[ P(\gamma = 1/\vec{\beta}, \vec{x}) = \frac{\exp(\vec{\beta}^T \vec{x})}{1 + \exp(\vec{\beta}^T \vec{x})} \]  

\[ P(\gamma = 0/\vec{\beta}, \vec{x}) = \frac{1}{1 + \exp(\vec{\beta}^T \vec{x})} \]  

Where \( x \) is a feature vector and \( \beta \) is a coefficient vector by using equation (3.4) we also have:

\[ P(\gamma = 1/\vec{\beta}, \vec{x}) = 1 - P(\gamma = 0/\vec{\beta}, \vec{x}) \]  

(6)

This form:

\[ P(\gamma = \pm /\vec{\beta}, \vec{x}) = \frac{1}{1 + \exp(\vec{\beta}^T \vec{x})} \]  

(7)

Where \( g \in \{\pm 1\} \) is the label for data item \( x \). It is also easily to verify that:

\[ P(g = 1/\vec{\beta}, \vec{x}) = 1 - P(g = -1/\vec{\beta}, \vec{x}) \]  

The Equivalence of Two Forms of Logistic Regression

At first glance, the form (3.4) and the form (3.6) look seem very different. However, the equivalence for these two forms can easily be established. Starting from the form (3.4), we can have:

\[ P(\gamma = 1/\vec{\beta}, \vec{x}) = \frac{\exp(\vec{\beta}^T \vec{x})}{1 + \exp(\vec{\beta}^T \vec{x})} \]  

(8)

\[ P(\gamma = 1/\vec{\beta}, \vec{x}) = \frac{1}{(1/\exp(\vec{\beta}^T \vec{x})) + 1} \]  

\[ P(\gamma = 1/\vec{\beta}, \vec{x}) = \frac{1}{(\exp(-\vec{\beta}^T \vec{x})) + 1} \]
Another way to establish the equivalent is from the classification rule. For the form (3.4), we have the following rule:

\[
P(y = 1 / \bar{\beta} x) = P(g = -1 / \bar{\beta} x) \]

(9)

\[
\frac{P(y = 1 / \bar{\beta} x)}{P(y = 0 / \bar{\beta} x)} > 1 \rightarrow y = 1
\]

(10)

\[
\frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)} > 1 \rightarrow y = 1
\]

(11)

**Applying the Model for Classification**

We consider two classes in this research, the presence of cancer among youth (label 1) and the absence of cancer among youth (label 0). The class decision is based on the result from 3.9 in the previous section. That is.

\[
P(y = 1 / \bar{\beta} x) = P(y = 0 / \bar{\beta} x)
\]

(9)

The \(P(y = 1 / \bar{\beta} x)\) is the probability of presence of the characteristic of interest (cancer among youth), while \(P(y = 1 / \bar{\beta} x)\) is the probability of absence of the characteristic of interest (cancer among youth). The quotient can be called an odd ratio.

**Principal Component Analysis Concept**

The PCA model is formulated as follows:

\[
X = TP + E
\]

where:

- \(X\) is the data matrix, with dimensions \(n \times p\), where \(n\) is the number of observations and \(p\) is the number of variables.
- \(T\) is the matrix of principal component loadings, with dimensions \(p \times k\), where \(k\) is the number of principal components.
- \(P\) is the matrix of principal component scores, with dimensions \(n \times k\).
- \(E\) is the matrix of residuals, with dimensions \(n \times p\).
The matrix $T$ is calculated by performing singular value decomposition (SVD) on the data matrix $X$. The $k$ columns of $T$ correspond to the $k$ largest singular values of $X$, and the $k$ rows of $T$ correspond to the $k$ principal components.

**Assumptions of Principal Component Analysis**

1. Linearity: this assumes the data set is a linear combination of the variables.
2. Importance of covariance and mean: There is no guarantee that the direction of maximum variance provides good features for discrimination.
3. Also, large variances have important dynamics: here, it assumes that components with larger variance correspond to interesting dynamics and lower ones correspond to noise.

**Research Findings**

The summary of the research findings was tabulated and presented below.

<table>
<thead>
<tr>
<th>Authors</th>
<th>The Research Focus</th>
<th>Design Methodology</th>
<th>Findings</th>
<th>Gaps Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asma Abdullahi et al</td>
<td>Predict cancer risk among adults using</td>
<td>Scoping review using the population, concept, and context approach. The study employs a quantitative research methodology</td>
<td>Identification of significant risk factors associated with cancer development among young adults, as determined through PCA and logistic regression analysis</td>
<td>Current understanding or literature regarding cancer risk factors among young adults, potentially paving the way for further research in this area to address these gaps and improve preventive strategies.</td>
</tr>
<tr>
<td>Wan Muhamad Amir et al</td>
<td>This study aims to develop a logistic regression model to predict the presence of oral squamous cell carcinoma (OSCC) using clinical and microbiological data.</td>
<td>The study employs a retrospective case-control design, analyzing data from patients diagnosed with OSCC and healthy controls. They use logistic regression to identify potential risk factors for OSCC based on the collected data.</td>
<td>His studies identify significant associations between several factors, including age, smoking status, alcohol consumption, and specific bacterial strains, and the risk of developing OSCC.</td>
<td>The authors acknowledge limitations such as the relatively small sample size and the single-center design, necessitating further validation in larger and diverse populations. Additionally, the study focuses on specific risk factors, and other potential contributors to OSCC might exist.</td>
</tr>
<tr>
<td>Liu et al.</td>
<td>This study investigates the effectiveness of machine learning models in predicting the development of bone metastasis in patients with advanced prostate cancer.</td>
<td>The study utilizes a retrospective analysis approach.</td>
<td>The study reports that the machine learning models achieved promising results in predicting bone metastasis. They found that the models could effectively distinguish patients with a high risk of developing bone metastasis from those with a low risk.</td>
<td>The authors acknowledge limitations in their research:</td>
</tr>
</tbody>
</table>

**Conclusion**

This research underscores the significance of employing statistical modeling methodologies, particularly logistic regression, and principal component analysis (PCA), in confronting the multifaceted challenges of cancer among the youth population in Nigeria, with a specific focus on regions such as Abuja and Kano. Through the utilization of these analytical frameworks, researchers can delve into intricate data patterns associated with cancer incidence, risk determinants, and prognostic outcomes. This analytical approach holds promise in facilitating early detection initiatives, devising tailored prevention strategies, and implementing timely interventions, all of which are attuned to the unique socio-economic and cultural dynamics prevalent in the region. Furthermore, this study has undertaken a comprehensive exploration of diverse statistical modeling techniques, underscoring the significance of comparative analyses and classification of prevalent cancer types affecting young individuals. Notably, logistic regression and PCA have been instrumental in this investigation. The development of predictive models for assessing risks and detecting these cancer variants at early stages stands poised to significantly augment healthcare outcomes for Nigeria’s youthful demographic. Leveraging logistic regression techniques can contribute to heightened precision in cancer prognostication, thereby enhancing the efficacy of preventive measures and clinical interventions.

**Recommendations**

- Future research endeavours should prioritize the validation and testing of statistical models in real-world settings within Nigeria. This approach ensures the applicability and reliability of these models in addressing the country’s specific healthcare challenges.
- Collaborative efforts among healthcare institutions, research organizations, and government agencies are imperative for the successful integration of statistical
models into clinical practice. Such partnerships facilitate the exchange of knowledge, resources, and expertise necessary for effective implementation.

- Increased investment in cancer research infrastructure, capacity building, and data collection efforts is essential. This investment will support ongoing initiatives aimed at addressing the escalating cancer burden among youth in Nigeria.

- Utilising statistical modeling techniques such as logistic regression and Principal Component Analysis (PCA) holds promise in advancing our understanding of cancer epidemiology. These techniques can also enhance early detection and diagnosis capabilities, ultimately contributing to the mitigation of cancer's impact on the youth population in Abuja, Kano, and other regions of Nigeria.

References
Glati, S., Cevivi, C. & Bernuzzi, A. (2007),” Symtomatic and Asymtomatic In Two Herosexual Couples,”
Journal of tropical medicine and Parasitology. 93(8)829-834.
