MODELLING HEALTH MAINTENANCE ORGANIZATIONS PAYMENTS UNDER THE NATIONAL HEALTH INSURANCE SCHEME IN NIGERIA

***Akinyemi M.I[1], Adeleke I.[2], Adedoyin C.[3]**
[1]Department of Mathematics, University of Lagos, Akoka, Lagos, Nigeria.
makinyemi@unilag.edu.ng
[2]Department of Business Administration University of Lagos, Akoka, Lagos,
adeleke22000@gmail.com
[3]Stamford University, Birmingham, Alabama, USA
christsonade@gmail.com

Abstract: *The Nigerian National Health Insurance Scheme (NHIS) is set up to ensure equitable payment of health care bills combining and prudently reducing cost-burden distribution for residents, versus high health care costs. Health maintenance organisations (HMO) are limited liability companies which could be established by private, public or individual entities with a main aim of being players in the scheme. This paper explored logistic regression (LR), linear discriminant analysis (LDA) and random forest (RF) in determining the factors that could determine if a HMO will cover full or part of an individual's healthcare bill. The results do not show a significant difference in the classification accuracies of the three methods. Inferring that the highest number of the Nigerian residents that make use of the NHIS lie between the 31-40yrs age bracket and that largely, ailment classification and the insured's age are key determining factors of whether a HMO would cover all or part of the bill.*

Keywords: *Random forests, Linear discriminant analysis, Logistic regression, Confusion matrix, Health care insurance*

1. INTRODUCTION

Increasing health care expenditures have given rise to numerous studies on the determinants of health care expenditures, health care policies, health care insurance, health care financing and other aspects of health care. In their 2006 report WHO, (2006), the public spending per capita for health in Nigeria was put at is less than $5, and dropping below $2 in some poorer parts of the nation. The Nigerian government thus inaugurated a committee in the National Council on Health which recommended the need for Health Insurance in Nigeria. In order to ensure the reach of the healthcare scheme, the NIHS put together different programmes aimed at accommodating diverse facets of the Nigerian sociocultural makeup. Modelling health insurance is an ongoing source of research; this is in a bid to deliver free, fair and accessible health care to the critical masses in the world economies.

2. LITERATURE REVIEW

Karanikolos, et al., (2013) in their paper study the effects recent economic crisis in Europe and the corresponding responses of governments has had on health systems. Ifanti, Argyriou, Kalofonou, & Kalofonos, (2013) explore data pertaining to effect of financial crisis as it relates to ascetical steps on health care, social services and health furtherance policies in Greece. Adeleke, Hamadu, & Ibiwoye, (2012) evaluates the NHIS capitation governance. Hamadu & Adeleke, (2012) built a model-assisted credibility assessment score for health Insurance claims in Nigeria. Ibiwoye & Adeleke, (2008) apply logistic regression to assess level of employee participation and observed that

spatial awareness of the scheme is a key factor affecting participation. However, it has been noticed that most times the health care organizations do not cater for the entire cost of treatment. It has been observed that often the insured person still must foot part of the medical bill. This study explores three predictive models viz: logistic regression (LR), linear discriminant analysis (LDA) and Random forest (RF) to identify the factors that could affect whether a HMO will cover full or a fraction of a person's healthcare bill. This study is very useful as this will give insights into the key factors that contribute to the kind of insurance coverage an individual can have access to.

The proceeding parts of this paper are set up as follows: In Section 3 a very brief overview of the methods employed in the research is given, we discuss the data as well as the results of some empirical analysis in Section 4. Finally, Section 5 concludes.

3. METHODOLOGY

3.1 Logistic Regression

The logistic regression response variable is usually dichotomous Pampel, (2000), that is, it can take the value 1 (probability, π), or 0 (probability,1-π). The model, given as: $P(Y) = \frac{1}{1+e^{-(b_0+b_1X_{1i}+\cdots+b_nX_{ni})}}$, P(Y) being the probability that Y will occur, b_0 is a constant, $X_{1i},\ldots,X_{ni}$ are the predictor variables and each b_i are the coefficients or weights attached to each predictor.

3.2 Random Forests

Random forest (RF) technique developed by Breiman, (2001) is based on the use of classification and regression trees. The RF classification procedure was carried out with the **randomForest** package in R see, Liaw & Wiener, (2002).

3.3 Linear discriminant analysis

In LDA, a linear combination of auxiliary variables is identified which maximises separation between categorical response groups Hastie, Tibshirani, & Friedman, (2009): $w_1 = a_1x_1 + a_2x_2 + \cdots + a_kx_k = \sum_{i=1}^{k} a_ix_i$. The weights a_i are chosen to maximize the separation between groups. For LDA the covariates are assumed to follow a multivariate normal distribution.

3.4 Confusion matrix

The performance of the classification system is evaluated using a confusion matrix Kohavi & Provost, (1998). The matrix consists of actual and predicted classifications. Table 1 gives the confusion matrix for a two-class classification. The cell entries in Table 1 in the context of our study mean the following:

Table 1: Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	n_{11}	n_{12}
	Positive	n_{21}	n_{22}

- a11 is the *specificity* i.e. the count of true predictions of negative instances (number of *True negatives (TN)),* calculated as : $TN = \frac{n_{11}}{n_{11}+n_{12}}$.
- a12 is the number of incorrect predictions that an instance is positive i.e. the number of

False positives (FP), calculated as: $FP = \frac{n_{12}}{n_{11}+n_{12}}$.

- a21 is the number of incorrect predictions that an instance negative i.e. the number of *False negatives (FN)*, calculated as: $FN = \frac{n_{21}}{n_{21}+n_{22}}$.
- a22 gives the *sensitivity* i.e. the number of correct predictions that an instance is positive i.e. the number of *True positives (TP),* calculated as: $TP = \frac{n_{22}}{n_{21}+n_{22}}$.

The overall accuracy (AC) calculated as: $AC = \frac{n_{11}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}$ is the proportion of the total number of predictions that were correct.

4. RESULTS

4.1 Data collection and data management

The data considered was collected from the Nigerian National Health Insurance Scheme. The data consists of over 55,000 cases spanning a 2-year period including variables age, classification of illness, kind of ailment, medical bill and HMO payment. Difference between medical bill and HMO payment was taken and then coded as follows: Yes, full cost of treatment catered for = 1, No, full cost of payment not catered for = 0.
Table 2 given below shows the distribution of patients who had their medical bills reimbursed in full or not.

Table 2: Frequency distribution of medical bill cover

Medical bill cover	Frequency	Percentage (%)
Medical bill not paid in full by HMO	38717	65.05
Medical bill paid in full by HMO	20801	34.95

Table 2 given above shows the frequency distribution of the insured who had their medical bill covered and those that did not. We observed here that most of the insured (65.05%) did not have their medical bill paid in full.

4.2 Logistic Regression

The logistic regression shows (Table 3) that both the age and the classification of ailment a are quite useful in determining whether an insured person will have their medical bill paid in full or not.

Table 3: Logistic regression coefficients

Parmeter	Estimate	Std. Error	P-value
(Intercept)	-1.721e+03	4.853e+01	<2e-16***
Age bands	4.851e-02	6.338e-03	1.95e-14***
Classification	4.591e-01	1.274e-01	0.000315***
Year	8.549e-01	2.411e-02	<2e-16***
(*** indicates significance at 99.9% confidence level)			

Furthermore, we computed the AUC of our model to be 61.23% i.e. given a random patient for whom the HMO pays in full, and a random

patient for whom the HMO pays part, our logistic model will correctly classify which is which about 61% of the time.

4.3 Confusion matrix

Table 4 given below presents the sensitivity and specificity and overall accuracy of the models

Table 4: Accuracy, Sensitivity, specificity and AUC

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC (%)
Logistic Regression (t= 0.5)	16.59	92.7	66.12	61.23
Random Forest	5.52	93.21	66.3	58.59
LDA	17.2	92.1	54.65	61.23

Table 4 given above indicates that all three models will have the similar specificity, that is they will all correctly classify the cases where the HMO will not pay the bill in full over 90% of the time.

5. Discussion & Recommendations

Analysis revealed that apart from the general and unspecified ailments (41.66%), pregnancy (9.14%), respiratory (12.54%) and digestive system (9.94%) related ailments make up for the bulk of the insured people while blood, ear, male genital, neurological, psychological and social problems related ailments each make up less than 1% of the insured. However, only the Urological ailment classification has more people with their bills fully paid by the HMO (0.83% paid as against 0.79% not paid). Predominantly bills are usually not fully paid by the HMO across all ailments.

In addition, we observed that most of the insured people belong to the 31-40 age group (41.04%) after which we have children less than 10years of age (27.06%) This amounts for the people who are most likely to be in paid employment at "white collar" jobs and their children. These people are likely to have access to the HMO services that their company offers. Furthermore, we found that in year 2013, an almost even split between those who had their medical bills fully paid and those who did not (12.60% vs. 12.44%). Furthermore, we also observed from **Error! Reference source not found.** that there was a rise in users of the health insurance scheme (i.e. from 1.11% to 73.84%) this figured dipped significantly in 2013 (25.04%).

Although the sensitivity of the models is very low the LR model and the LDA model seem to have a better sensitivity than the RF model as they will correctly classify the number of cases where the HMO will pay at the medical bill in full about 16.59% and 17.2 % of the time respectively while the random forest model will only classify such about 5.52% of the time. On the other hand, the overall accuracy of the LR models and the Random forest are not far apart hovering around 66% while that of the LDA model is just over 50%.

It is recommended that the government generate more awareness of the NHIS to other clusters of the population especially the unemployed and underemployed members of the society. Furthermore the older dependent members of the populations should be encouraged to take

advantage of the scheme.

6. Conclusion

The distribution of the medical bill cover of the people insured under the Nigerian NHIS system was studied. The instance of medical bill payment by the HMOs was modelled using a logistic regression model with age bands, ailment classification and year of claim as the variables. Comparison was made between the logistic regression model, the random forests and linear discriminant analysis. It was found both age, and ailment classification are useful in predicting whether and insured person's medical bill will be paid in part or in full. It was also found that most of the beneficiaries of the NHIS system are the working class of the population and their children.

Furthermore, the model comparison shows that the logistic regression and the random forests both have relatively better accuracy than the linear discriminant analysis models, all three models have similar strong specificity, however, the Random forest has the lowest sensitivity and the linear discriminant model has the highest specificity.

References

Adeleke, I., Hamadu, D., & Ibiwoye, A. (2012). Evaluation of the capitation regime of Nigeria Health Insurance Scheme. *International Journal of Academic Research Part A, 4*(5).

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Hamadu, D., & Adeleke, I. (2012). Model-assisted credibility rating for health Insurance claims. *Journal of Mathematics and Technology, 3*(2), 32-37.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). THE ELEMENTS OF STATISTICAL LEARNING. In T. Hastie, R. Tibshirani, & J. Friedman, *Springer Series in Statistics.* Springer.

Ibiwoye, A., & Adeleke, I. A. (2008). Does National Health Insurance Promote Access to Quality Health Care? Evidence from Nigeria. *The Geneva Papers on Risk and Insurance - Issues and Practice, 33*(2), 219-233.

Ifanti, A. A., Argyriou, A. A., Kalofonou, F. H., & Kalofonos, H. P. (2013). Financial crisis and austerity measures in Greece: Their impact on health promotion policies and public health care. *Health Policy, 113*(1-2), 8-12.

Karanikolos, M., Mladovsky, P., Cylus, J., Thomson, S., Basu, S., Stuckler, D., . . . McKee, M. (2013). Financial crisis, austerity, and health in Europe. *The Lancet, 381*(9874), 1323-1331.

Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. *In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, 30*.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Pampel, F. (2000). Logistic regression: a primer. In F. Pampel, *Sage university papers series: Quantitative applications in the social sciences.* Sage Publications.