

On Statistical Analysis of Fraud Risk and Preventive Care as Determinants of Claim Severity in the Kenyan Retail Health Insurance

Shelmith V. M. Mwangi, Richard Simwa and Michael kirumbu

Daystar University, Nairobi, Kenya

DOI: <https://dx.doi.org/10.47772/IJRISS.2025.909000818>

Received: 12 September 2025; Accepted: 18 September 2025; Published: 31 October 2025

ABSTRACT

Retail health insurance in Kenya faces growing challenges from rising healthcare costs, fraudulent claims, and limited adoption of preventive care. These factors increase premiums, reduce affordability, and hinder private insurance uptake compared to public schemes. This study evaluates the impact of fraud and preventive healthcare on claim severity, alongside demographic and behavioral influences. Using Pearson correlation and Generalized Linear Models (GLM), relationships between fraud incidence, preventive care participation, and selected demographics were analyzed, with significance tested through p-values and R-squared metrics. Findings reveal that fraud markedly raises claim severity, inflating costs and undermining sustainability, while preventive care reduces long-term claim costs. Demographic and behavioral factors show some influence on claim behavior but remain less significant than fraud and preventive care. The study concludes that incorporating fraud detection and preventive healthcare into actuarial pricing models can enhance affordability and sustainability of retail health insurance in Kenya, providing valuable guidance for insurers and policymakers.

Keywords: Retail health insurance, Fraudulent claims, Preventive healthcare, Claim severity, Actuarial models, Insurance pricing, Healthcare cost

INTRODUCTION

Retail health insurance in Kenya is critical for expanding healthcare access, particularly for populations outside employer-sponsored schemes. However, the sector faces persistent challenges, including high premiums, rising claim costs, and low uptake relative to public health financing. Insurance fraud, ranging from inflated hospital bills to collusion and false diagnoses, drives claim severity and undermines trust.

Preventive healthcare remains underutilized, with most products focusing on treatment rather than proactive risk reduction. While existing studies focus on access and affordability, few examine how fraud and preventive healthcare interact to influence claim severity from an actuarial perspective. This study addresses this gap by analyzing fraud risk, preventive care participation, and demographic influences, aiming to provide actionable insights for sustainable, data-driven insurance practices in Kenya.

LITERATURE REVIEW

In order to investigate fraud and its prevention in Healthcare Insurance, we note the following literature.

Fraudulent claims inflate healthcare costs globally, often accounting for 10-25% of payouts (Coalition Against Insurance Fraud, 2022). In Kenya, the retail health insurance segment is prone to upcoding, phantom billing, and collusion (IRA, 2023; AKI, 2020). Parliamentary investigations into the NHIF (now SHA) revealed systemic

fraud in claims processing (Parliament of Kenya, 2024). Studies highlight the growing use of machine learning and data mining for fraud detection (Moturi, 2019; Muthura, 2024), but actuarial models rarely incorporate fraud directly into pricing frameworks, leaving premiums vulnerable to inflation (Wüthrich & Merz, 2008).

Preventive care, including screenings, vaccinations, and wellness programs, reduces long-term claim costs and strengthens risk pools (Dobson & Barnett, 2018; WHO, 2022). In Kenya, preventive measures are underutilized despite policy support (SHA, 2022). Empirical studies linking preventive-care participation to claim severity in actuarial models are limited.

GLMs are widely used to model insurance claim frequency and severity, offering interpretability, regulatory alignment, and inferential insights (Dobson & Barnett, 2018; Masese, 2020). While machine learning can outperform GLMs in prediction, GLMs remain practical for actuarial pricing. Prior Kenyan studies show feasibility, though challenges such as imbalanced fraud labels and limited preventive-care data persist (Moturi, 2019; Owuor, 2023).

Contribution: The study develops GLM-based models integrating fraud and preventive-care variables, shifting focus from detection to estimating causal effects on claim severity, providing actionable insights for pricing and policy design.

METHODOLOGY

The study employed a quantitative research design supported by actuarial and statistical modeling to analyze the effects of health insurance fraud and preventive care on claim severity in Kenya's retail health insurance market. The methodology covers data preparation, severity simulation, regression modeling, and statistical testing.

Data Collection and Preparation

A simulated dataset of 31 individuals was created to reflect real-world retail health insurance claims. Variables included:

Age (continuous)

Fraud Risk Level (categorical: High, Moderate, Low)

Preventive Care Participation (Yes/No)

Region (Rural/Urban)

Education Level (Primary, Secondary, University)

Claim Severity (continuous, dependent variable)

Categorical variables were transformed into dummy variables for regression and correlation analysis. Fraud Risk Level was converted into a Fraud Factor, with values assigned based on literature estimates (High = 1.20, Moderate = 1.10, Low = 1.00).

Simulated Severity Calculation

Claim severity was generated using a risk-adjusted actuarial model:

$$\text{Severity} = C \times F \times (1 - \delta * P) + \varepsilon \quad (1)$$

Where:

C = Base claim cost (KES 10,000)

F = Fraud factor

δ = Preventive care discount rate (0.15)

P = Preventive care usage (1 = Yes, 0 = No)

ε = Random error term

This formulation reflects actuarial frameworks for incorporating fraud risk and preventive healthcare into claim cost estimates (Derrig, 2002; Esmaili & Deng, 2017).

Regression Modeling Using GLM

A Generalized Linear Model (GLM) was developed to quantify the effects of demographic, fraud, and preventive care variables on claim severity:

$$\text{Severity} = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Fraud Factor} + \beta_3 \cdot \text{Preventive Care} + \beta_4 \cdot \text{Region Dummies} + \beta_5 \cdot \text{Education Dummies} + \varepsilon \quad (2)$$

The GLM was chosen for its ability to handle mixed data types and skewed insurance claim distributions. Coefficients were estimated using Ordinary Least Squares (OLS), with t-tests and p-values used to assess statistical significance. The model's explanatory power was evaluated using the R-squared statistic.

Statistical Techniques

Correlation Analysis

Pearson's correlation coefficient was used to measure the strength and direction of relationships:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (3)$$

Significance Testing

A t-test was applied to assess statistical significance where

$$t = r(n-2)^{1/2} / (1-r^2)^{1/2} \quad (4)$$

The computed p-value then leads to acceptance or rejection of the corresponding hypothesis.

Coefficient of Determination (R^2)

To evaluate the proportion of variance in claim severity explained by fraud and preventive healthcare measures, the R^2 statistic was computed as:

$$R^2 = 1 - (SS_{\text{Regression}}) / (SS_{\text{Total}}) \quad (5)$$

Data preparation and initial calculations were carried out using Microsoft Excel, including creation of dummy variables and descriptive summaries. Statistical formulas were applied to compute correlations, t-values, and p-values, allowing for the assessment of relationships between key variables. An excerpt of claim severity calculations is shown in Appendix A.

RESULTS

Application of the regression model given in Equation [2] and the correlation tests indicated by Equation [3] and [4] led to the results in Table 1 and Table 2 respectively.

Table 1 : Statistical Tests for Regression

Explanatory Variable	Regression coefficient	p-value
Intercept	398.25	-
Region	0.1942	0.2985
Education Level [secondary]	-0.24	0.1936
Education Level [university]	0.3249	0.0747
Age	-0.0217	0.9890
Has Prevention	-0.7913	0.0000*
Fraud Factor	0.7667	0.0000*

Fraud Factor demonstrated a statistically significant positive correlation with claim severity. However, Preventive care participation showed a statistically significant negative correlation with claim severity. Overall, education showed no statistically significant relationship with claim severity, though university-level education may indicate a trend toward higher severity. Age was positively correlated with claim severity but statistically not significant.

Table 2 : Statistical Tests for Correlation

Variable	r	t	p-value
Fraud	0.7667	6.43	0.00001*
Preventive care participation	-0.7913	-6.97	0.000001*
Age	0.3281	1.89	0.069
Secondary Education	-0.2400		0.1932
University Education	0.3249		0.0746

Fraud Factor demonstrated a statistically significant strong positive correlation with claim severity. Preventive care participation also showed a statistically significant but strong negative correlation with claim severity. Age was not statistically significant though it showed a weakly positive correlation with claim severity. Secondary education had weak negative correlation which was not statistically significant. University education had moderate positive correlation which was not statistically significant.

The causes of variation in the claim severity was analyzed using the R² statistic given in Equation [5]. The model achieved an R² value of 0.72, indicating that 72% of the variation in claim severity was explained by mainly the two variables; fraud factor and preventive care participation.

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

The findings indicate that fraudulent claims and preventive healthcare have a significant impact on the severity of insurance claims, while demographic variables such as age and gender showed limited explanatory power. The positive correlation between fraud and claim severity highlights the need for stronger fraud management mechanisms, as unchecked fraudulent activity inflates costs and directly undermines affordability. Conversely, the negative association between preventive healthcare and claim severity demonstrates its potential as a cost-containment measure. These insights suggest that insurers should reframe their risk modeling approaches to place greater emphasis on behavioral and risk-related factors rather than over-relying on demographic predictors.

This study concludes that fraud remains a critical driver of rising claims severity in Kenya's retail health insurance sector, while preventive healthcare offers a practical pathway for reducing costs and enhancing sustainability. Demographic factors such as age and gender, although traditionally used in underwriting, were not statistically significant in explaining claim severity. Therefore, insurance providers and policymakers must reconsider how they structure health insurance products, shifting focus toward proactive risk management strategies.

Drawing from the results, several recommendations are proposed:

Strengthen Fraud Detection. Insurers should invest in advanced fraud detection systems, including AI-driven anomaly detection and predictive analytics, to curb the escalating impact of fraudulent claims.

Incentivize Preventive Healthcare. Insurance products should embed preventive health benefits and incentives, such as premium discounts, wellness programs, and rewards, to encourage uptake and reduce claims over the long term.

Refocus Risk Modeling. Actuarial models should place more weight on behavioral and risk-linked variables, particularly fraud and preventive healthcare, which demonstrated stronger predictive power than demographics.

Policy Support for Preventive Health. Regulators should align with insurers to promote preventive healthcare initiatives within the framework of universal health coverage (UHC), thereby enhancing both affordability and access.

Future Research. Further studies using actual claims datasets and advanced actuarial/statistical techniques (e.g., R-based modeling, machine learning) are recommended to validate these findings and provide stronger evidence for industry-wide adoption.

REFERENCES

1. Association of Kenya Insurers. (2020). Insurance industry annual report. Nairobi, Kenya: AKI.
2. Coalition Against Insurance Fraud. (2022). The impact of fraud on the insurance industry. Washington, DC: CAIF. <https://insurancefraud.org>
3. Derrig, R. A. (2002). Insurance fraud. *Journal of Risk and Insurance*, 69(3), 271-287. <https://doi.org/10.1111/1539-6975.00023>
4. Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models (4th ed.). CRC Press. <https://doi.org/10.1201/9781315182780>
5. Field, E. (2018). Preventive healthcare incentives in insurance: Evidence from developing countries. *Journal of Health Economics*, 62, 1-14. <https://doi.org/10.1016/j.jhealeco.2018.07.002>
6. Insurance Regulatory Authority. (2023). Annual insurance industry report 2022. Nairobi, Kenya: IRA. <https://www.ira.go.ke>
7. Maina, A. K. (2024). Fraud management strategies and performance of medical insurance providers in Nairobi City County [Master's thesis, University of Nairobi]. University of Nairobi Digital Repository.
8. Parliament of Kenya. (2024). Report of the inquiry into alleged fraudulent payments of medical claims and capitation by NHIF. Nairobi, Kenya: Government Printer.
9. Social Health Authority. (2022). Strategic plan 2022-2027. Nairobi, Kenya: SHA.
10. World Health Organization. (2022). Preventive care and universal health coverage. Geneva, Switzerland: WHO. <https://www.who.int>