Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 18, No. 2, 2025

An Empirical Analysis of The Influencing Factors of America's Medical Insurance Cost

Tianyi Deng *, Shuai Gong, Wanlu Li, Tairan Zhou

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University (XJTLU), China
* **Corresponding author**: Tianyi Deng (Email: Tianyi.Deng22@student.xjtlu.edu.cn)

Abstract: This paper empirically analyzes the factors influencing the cost of medical insurance in the United States. Based on data from 12,761 American individuals sourced from Kaggle, the study examines the relationships between age, gender, body mass index (BMI), diabetes, smoking status, and medical insurance costs. Multiple linear regression models were employed, and model optimization techniques, such as log-linear transformation, were applied to enhance the explanatory power of the analysis. The results indicate that age, BMI, diabetes, and smoking status have significant positive correlations with medical insurance costs, while gender (female) shows a negative correlation. The final log-linear model achieved a high goodness-of-fit (R^2=0.694), demonstrating that these factors effectively explain variations in medical insurance costs. The study provides valuable insights for insurance companies in premium pricing and for individuals in health management.

Keywords: Medical insurance cost, BMI, Smoking status, Multiple linear regression, log-linear model.

1. Motivation

Health insurance covers medical expenses incurred by individuals. It is a vital tool for individuals to manage the high expenditure of healthcare, and it is essential to ensure that people of all income levels have access to medical treatment. Ramachandran, Kavitha and Pandimeena (2023) emphasized that the effectiveness and provision of healthcare services must be improved to increase the productivity of human capital. World Health Organization (2021) reported that the cost of healthcare has been steadily rising worldwide between 2000 and 2018, accounting for almost 10% of global GDP. Moreover, Empirical data points to a link between insurance and economic expansion (Singhal, Goyal and Singhal, 2022). Consequently, given the social and economic significance of health insurance premiums, analyzing the factors that affect healthcare insurance costs is essential.

Orji and Ukwandu (2023) claimed that this is essential for both luring and maintaining policyholders and effectively managing current insured. Indeed, actuarial modeling of insurance costs has emerged as a major field of study in the medical insurance industry in recent years (Duncan et al., 2016). Zhang (2019) employed a mediated effects model to investigate the relationship between environmental pollution and health insurance expenditures, revealing that environmental pollution substantially elevates medical insurance costs. He (2024) adopted the multiple linear regression model and the stepwise regression method, whose results reveal that age, BMI, number of children, smoker or non-smoker and region have plausible correlations with medical insurance cost. It is illustrated by Li (2021) that through the Dagum Gini coefficient method to a sample of 31 Chinese regions' statistics from 2010 to 2019, the amount spent on medical insurance funds varied by location. Using a logistic regression model, Zhou and Zhao (2022) demonstrated that customers who use the WeChat Healthcare pedometer function more frequently had lower rates of health insurance.

Accurately developing a forecast model for medical insurance costs is difficult, nevertheless, because of the multiplicity of elements and their complexity (Orji and Ukwandu, 2023). Therefore, this paper utilizes econometric models to quantitatively examine the impact of age, BMI, diabetes, smoker and gender on the cost of medical insurance claims. Based on the analysis of these determinants, the article offers recommendations to both insurance companies and policyholders, to enable patients to make more informed decisions during medical treatments while allowing insurance companies to better to charge each customer an appropriate premium for the risk they represent, aiming to support the sustainable development of the health insurance industry.

2. Data Description

2.1. Source of Data

This essay empirically explores the relationship between six variables and health insurance cost. Data was all aggregated from the Kaggle repository, which was collected by Suresh Gupta and contained 12761 entries from Americans. The 12761 individuals in the data used in this article all spent money on health insurance to some extent. They range in age from eighteen to sixty-four. The independent variables are age (A), gender (G), body mass index (BMI), diabetes (D), and smoker (S) and the dependent variable is medical insurance cost (MIC).

2.2. Description of the Variables

2.2.1. Medical insurance cost (MIC)

Plemans (2024) stated that health insurance claims are formal requests made by policyholders to insurance companies for reimbursement of medical services they have received. The ability to predict a correct claim amount which is considered as the medical insurance cost has a critical influence on insurers for considering fixing the health premium to better manage and allocate health insurance funds.

2.2.2. Age (A)

The risk of diagnosed diseases increases with age, older people usually tend to decline in physical function. Health insurance providers assess people based on age and offer health insurance premiums accordingly (Anderson et al.,

2012).

2.2.3. Body mass index (BMI)

BMI is defined as an indicator to measure obesity. A BMI over 24.9 is considered overweight and over 30 is considered obese. Xia (2023) revealed that, if a person's BMI is at a higher level, then that person will face a higher probability of transaction from health to death, and the cost of medical insurance claims is also higher.

2.2.4. Diabetes (D)

Diabetes is a typical chronic disease, and patients need to invest in long-term medical expenses to maintain relative health of the body system. In recent years, the number of people with diabetes has increased, and the economic cost has also increased, mainly because having diabetes makes health insurance more expensive for that person (American Diabetes Association, 2018).

2.2.5. Smoker (S)

According to epidemiological studies, smoking behavior will increase the probability of cancer, a serious threat to health (Yun et al., 2005). Consequently, smoking can raise the opportunity of developing certain medical disorders, and insurers may need to pay higher medical insurance costs for claims.

2.2.6. Gender (G)

Males and females also have different physical conditions which means that the probability of diseases might vary by gender. Agnieszka (2020) claimed that women have greater medical needs for parturition and gynecological problems, so women's health insurance costs are higher than men's.

Table 1. Description of the Variables

Variable	Description	Unit
Medical insurance cost (MIC)	Medical cost of the client per year. The number of claims that insurance companies need to pay for the policyholders.	US dollar
Age (A)	Present Age of the client	Year
Body mass index (BMI)	$BMI = \frac{weight}{height^2}$	kg/m^2
Diabetes (D)	No – 0, Yes – 1	None
Smoker (S)	No – 0, Yes – 1	None
Gender (G)	Male – 1, Female -0	None

Source: Author-compiled

2.3. Descriptive Statistics

The descriptive statistics for initial data and the relative.

Table 2. Descriptive Statistics for Initial Data

	MIC	A	BMI	D	S	G
Observation	12761	12761	12761	12761	12761	12761
Mean	11591.45	39.63	30.04	0.77	0.15	1
SD	10018.61	14.06	6.05	0.42	0.36	0.50
Minimum	1121.90	18	16.00	0	0	0
Median	8964.10	40	29	1	0	0
Maximum	58571.10	64	53.10	1	1	1

Source: Author-compiled

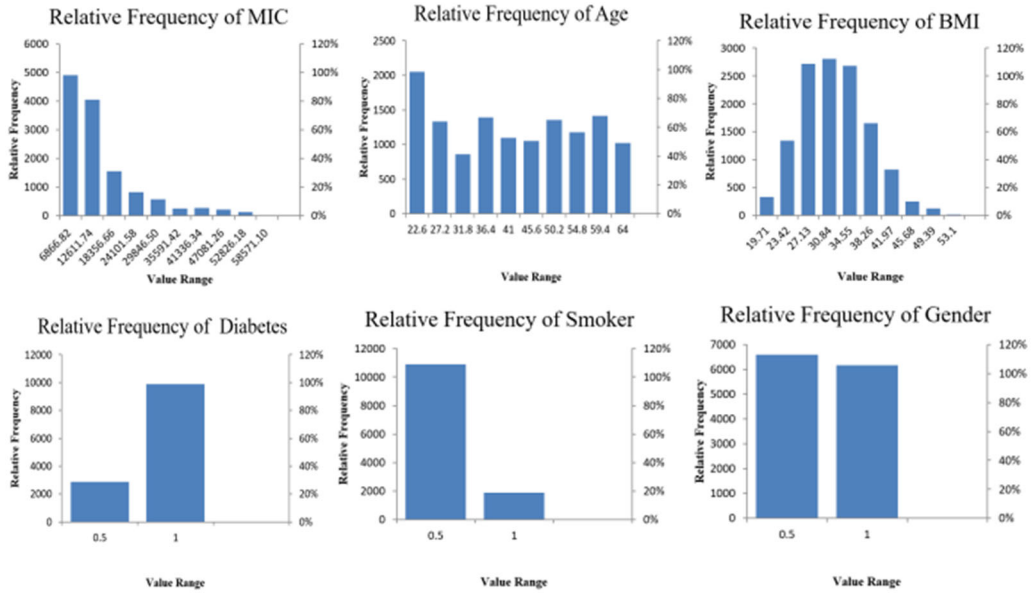

Figure 1. Relative Frequency of Variables

2.4. Data Processing

The correlation between the variables is shown in Table 3. All five variables have a positive correlation with MIC, particularly age and smoking. This enables us to confirm that every variable in the model is statistically significant and to infer the relationships between the variables in a preliminary manner.

More specifically, we examine the variance inflation factor (VIF) to test multicollinearity. Table 4 shows that the VIF values for all variables are close to 1, and the corresponding 1/VIF values range from 0.951 to 0.992, indicating that there

is no significant multicollinearity issue in the model. The variables are nearly independent, with a small linear correlation between them. This means each variable contributes independently to the model without interference, ensuring the model's stability. Therefore, there is no need to adjust the variables.

Table 3. Correlation

		MIC	A	BMI	D	S	G
MIC	Correlation	1					
	P-value						
A	Correlation	.377**	1				
	P-value	0.000					
BMI	Correlation	.137**	.192**	1			
	P-value	0.001	0.001				
D	Correlation	.091**	.069**	.063**	1		
	P-value	0.001	0.001	0.001			
S	Correlation	.694**	-.018*	-.061**	.012	1	
	P-value	0.000	0.001	0.001	0.175		
G	Correlation	.049**	.017	.068**	-.022*	.070**	1
	P-value	0.001	0.052	0.001	0.014	0.001	

*p<0.05, **p<0.01, ***p<0.001
Source: Author-compiled

Table 4. VIF

Variable	VIF	1/VIF
A	1.040	0.960
G	1.010	0.989
BMI	1.050	0.951
D	1.010	0.992
S	1.010	0.991
Mean VIF	1.020	

Source: Author-compiled

3. Empirical Model and Hypotheses

3.1. Hypothesis

H0 (Null Hypothesis): $\beta 1=0$

There is no significant relationship between the main variable and the dependent variable.

H1 (Alternative Hypothesis): $\beta 1\neq 0$

There is a significant relationship between the main variable and the dependent variable.

3.2. Hypothesis Testing for the Main Variable

Based on the hypothesis test results from Table 5, Age is highly suitable as the main variable due to its statistically significant and strong relationship with MIC. The P-value for age is 0.000, allowing us to reject the null hypothesis and conclude that age has a significant impact on the dependent variable. Furthermore, the robust standard error of 5.845 and a high t-test value of 45.91 confirm the precision and strength of this relationship. From a theoretical perspective, it is also reasonable to expect that age could influence the dependent variable, as aging may correlate with higher health risks or insurance claim amounts.

Table 5. Hypothesis test

MIC	Coefficient	Robust Stand Error	t-test	P-value
A	268.340	5.845	45.91	0.000
Constant	957.204	245.768	3.89	0.000

Source: Author-compiled

3.3. Model Selection

Based on these findings, choosing the multiple linear regression model aligns well with the basic principles of econometrics, which state that maintaining the validity and integrity of the model's estimations depends critically on assigning the variable of interest nearly at random while holding the control variables constant. This method improves the study's analytical rigor and the findings' dependability, allowing for an objective and reliable age coefficient estimator. Additionally, it guarantees that the cost of medical insurance by age is interpreted causally.

3.3.1. Model 1

According to the result of hypothesis testing, we set age as the main variable.

$$MIC=\beta_0+\beta_1 A+\beta_2 BMI+\beta_3 D+\beta_4 G+u_i,\ i=1,\ 2,\ 3...12761$$

MIC: dependent variable represents medical insurance cost

A, BMI, D, G: Age, Body Mass Index, Diabetes and Gender, respectively

u_i: regression error

β_0: population intercept

β_1, β_2, β_3, β_4: the corresponding coefficients for the independent variables

Table 6. Model 1 Result

Variable	Coefficient	Robust Stand error	t-Statistic	P-value
A	256.340***	6.099	42.03	0.000
BMI	101.323***	13.866	7.31	0.000
D	1521.672***	164.617	9.24	0.000
G	804.013***	164.303	4.89	0.000
Constant	-3177.339***	439.881	-7.22	0.000
R-squared 0.152				

Source: Author-compiled

The results are promising, with the p-values for all variables being 0.000, indicating that the coefficients in the model are statistically significant. However, the R-squared value is relatively low, so the model has limited explanatory power. We believe this may be due to the omission of relevant variables.

3.3.2. Model 2

To enhance the model's goodness-of-fit and account for omitted causal effects that could lead to omitted variable bias, we draw on theoretical knowledge to hypothesize that smoking status may influence the medical insurance costs incurred by the insured. This is based on the premise that smoking increases the likelihood of developing certain diseases, which in turn could result in higher claim amounts.

Furthermore, as observed in Table 3 and Table 4, although the correlation between Smoker and the existing independent variables is relatively weak, it is still meaningful from a statistical perspective. Furthermore, Smoker exhibits a strong and significant correlation with the dependent variable MIC. Therefore, the Smoker variable is introduced to attempt to reduce omitted variable bias and enhance the explanatory power of the model.

$$MIC=\beta_0+\beta_1A+\beta_2BMI+\beta_3D+\beta4G+\beta_5S+v_i,\ i=1,\ 2,\ 3...12761$$

S: represent Smoker
v_i: the error term in model 2
β_5: coefficient for Smoker

Table 7. Model 2 Result

Variable	Coefficient	Robust Stand error	t-Statistic	P-value
A	260.554***	3.875	67.25	0
BMI	178.595***	9.093	19.64	0
D	1210.529***	129.655	9.34	0
S	19988.93***	237.944	84.01	0
G	-248.301**	105.710	-2.41	0.016
Constant	-8374.527***	296.885	-26.48	0
R-squared 0.648				

Source: Author-compiled

After including the smoke variable, the R-squared value significantly improved and all variables remained statistically significant. This is a positive outcome, indicating that we have successfully identified and included an omitted variable.

However, it is curious that the coefficients in the new model have increased substantially, with the smoker's coefficients close to 20,000. Therefore, we began to explore the potential causes behind this issue.

3.3.3. Model 3

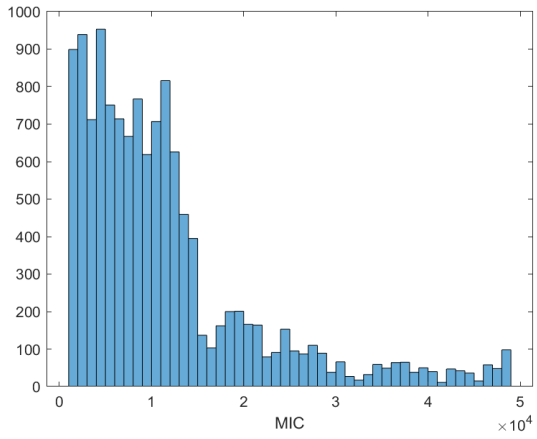

Figure 2. histogram of MIC

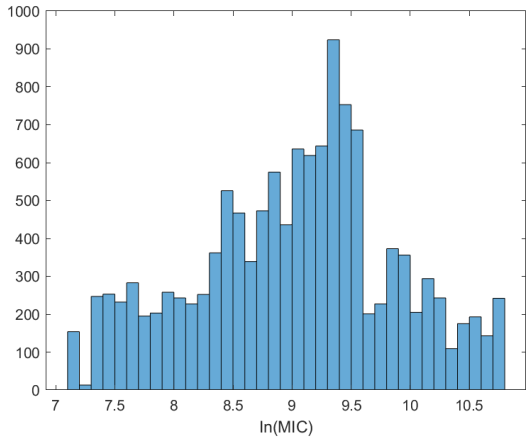

Figure 3. histogram of Ln (MIC)

When plotting the histogram of MIC, it becomes evident from Figure 2 that the majority of the data values are large, resulting in a right-skewed distribution. Right-skewed data typically contain extremely high values, which might inappropriately influence on the regression model, potentially leading to instability. According to theoretical principles, linear regression models assume that the error terms are normally distributed. However, skewed data generally fail to meet this assumption of normality.

Benoit (2011) emphasized that one practical way to make a severely skewed variable more roughly normal is to apply logarithmic transforms. Subsequently, we applied a logarithmic transformation to the dependent variable. As shown in Figure 3, the transformed data now approximates a normal distribution. Therefore, to solve the issue we found in model 2, we introduce log-linear.

$$Ln\ (MIC)=\beta_0+\beta_1A+\beta_2BMI+\beta_3D+\beta_4G+\beta_5S+vi,\ i=1,\ 2,\ 3...12761$$

Additionally, from the scatterplots, it can be observed that the model used in Figure 4 exhibits an uneven distribution of data points on either side of the fitting line, indicating the presence of bias or error patterns. In contrast, the model used in Figure 5 shows a more uniform distribution of data points around the fitting line, reflecting a more ideal-fitting performance with residuals that do not exhibit significant bias. Therefore, we have effectively applied the log-linear model.

Table 8. Model 3 Result

Variable	Coefficient	Robust Stand error	t-Statistic	P-value
A	0.036***	0.000	102.77	0.000
BMI	0.008***	0.001	12.31	0.000
D	0.075***	0.010	7.24	0.000
S	1.375***	0.013	108.68	0.000
G	-0.071***	0.008	-8.42	0.000
Constant	7.095***	0.023	307.95	0.000
R-squared 0.694				

Source: Author-compiled

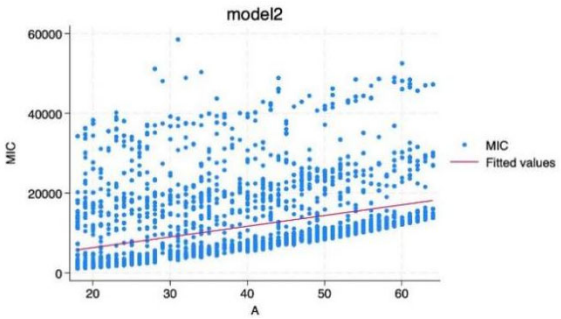

Figure 4. Scatter-plot of Model2

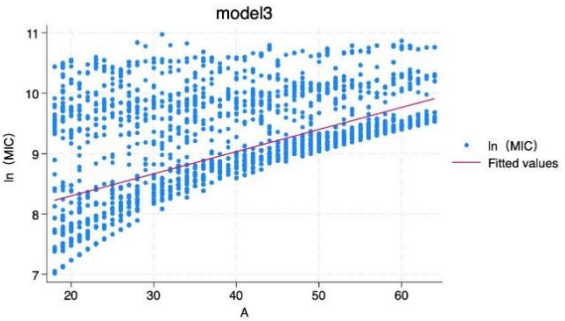

Figure 5. Scatter-plot of Model3

4. Regression Results and Inferences

Our model iteration effectively increases the reliability of the estimators during the process of updating the three models, assuming that the estimations of the coefficients of age stay almost constant. Additionally, the final log-linear level of fit has significantly increased, going from 0.152 in model 1 to 0.694 in model 3.

Observing more deeply into the coefficients from Table 8 reveals nuanced relationships between health insurance costs and personal characteristics. A and MIC have a positive link, and the coefficient is 0.036 ($p<0.001$) which indicates that people's medical needs typically increase as they age, leading to higher health insurance costs. For instance, older people may be more prone to chronic diseases that require more medical services and prescription drugs, so insurers need to charge higher premiums to cover those costs, considering the claim of healthcare insurance is higher. Additionally, BMI positively correlated with MIC (coefficient 0.008, $p<0.001$). An abnormal body mass index in real life will raise the incidence of certain illnesses, such as cardiovascular disease and diabetes. People with a higher BMI will pay higher premiums which are calculated by insurance providers because these health issues can result in more medical expenses and more insurance costs that insurers need to spend. Similarly, there is a positive correlation (0.075, $p<0.001$) between diabetes and medical insurance expenditures. Diabetes is a chronic disease that requires long-term treatment, including regular checkups, medication, and management of possible side effects. People with diabetes consequently pay more for healthcare. There is a high correlation between MIC and smoking behavior, as evidenced by the significant smoker coefficient (1.375, $p<0.001$). Smoking is a contributing factor to many hazardous diseases, including lung cancer and cardiovascular disease. Therefore, smokers might need more costly healthcare, insurance companies raise the premiums for smokers as they have a higher risk of paying more money on claims.

The cost of medical insurance is lower for males than for females, as indicated by the negative correlation of -0.071 ($p<0.001$) between gender and medical insurance expenses in Model 3. The potential explanation could be that women have particular health requirements. For example, the costs associated with childbearing are high and include pregnancy testing, delivery, postpartum rehabilitation, and several other procedures that necessitate ongoing medical resource investment. Moreover, certain gynecological conditions, such as breast diseases and uterine fibroids, are more prevalent. In terms of the awareness of preventive health care, females might take the initiative to perform routine physical examinations, screening various diseases, and engage in other preventive medical behavior, so the frequency of women's medical treatment is relatively high, and the overall medical cost may be higher, which makes the insurance cost of women relatively higher than man, and there is a negative correlation with the male sex.

5. Conclusion and Discussion

This essay analyzes the impact of age, BMI, diabetes, smoker and gender on the cost of medical insurance based on 12761 individuals from Americans. The study initially employed correlation analysis and VIF to assess the relationships between the independent variables and the dependent variable, as well as to identify whether there is potential multicollinearity among the factors. Subsequently, hypothesis testing was conducted on the primary variable, age, to validate its significant effect on the dependent variable. A multiple linear regression model was then applied, incorporating the control variable smoker to reduce omitted variable bias. Upon further analysis, it was observed that the dependent variable exhibited a right-skewed distribution with extreme values, which could compromise the stability of the model. To address this issue, the model was modified to a log-linear form. The results ultimately indicated that the log-linear model was the more suitable model for this study and these variables can be used to predict and explain medical insurance costs to a great extent.

Despite the progress achieved in our study, certain limitations remain. After three iterations of model refinement, the R-squared of the model reached only 0.694, indicating that the fit still requires further improvement. This may be attributed to the current selection of variables—our analysis considered only five variables with MIC, overlooking other influential factors such as regional environment and social economic conditions. Furthermore, future research could explore the interaction effects among variables or apply more sophisticated models to enhance the explanatory power and predictive performance of the analysis.

In addition, we have provided some practical recommendations for the two groups. For insurance companies, charging an appropriate premium based on the risk represented by each client is crucial. The ability to accurately predict claim amounts, or in other words, the health insurance costs, has a significant impact on the company's management decisions and financial statements. Based on this research, insurance companies could predict the medical insurance costs that may be incurred by the insured individual by taking into account factors such as age, gender, BMI, smoking status, and diabetes. This allows the company to ultimately determine the premium that the client should pay. On the other hand, for policyholders, it is evident that physiological factors have a considerable impact. Therefore, it is recommended that individuals maintain better health and

eliminate unhealthy habits in to reduce medical insurance costs and save money.

References

[1] American Diabetes Association (2018) 'Economic costs of diabetes in the U.S. in 2017', Diabetes Care, 41(5), pp. 917-928. Available at: https://doi.org/10.2337/dci18-0007 (Accessed: 12 November 2024).

[2] Anderson, M., Dobkin, C. and Gross, T. (2012) 'The Effect of Health Insurance Coverage on the Use of Medical Services', American Economic Journal: Economic Policy, 4(1), pp. 1–27. doi: 10.1257/pol.4.1.1

[3] Benoit, K. (2011) Linear Regression Models with Logarithmic Transformations London: London School of Economics, pp. 1-8.(Accessed: 12 November 2024).

[4] Duncan, I., Loginov, M., and Ludkovski, M. (2016) 'Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs', North American Actuarial Journal, 20(1), pp. 65-87.(Accessed: 12 November 2024).

[5] Gupta, S. (2022) Health insurance data set. Available at: https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set/data (Accessed: 10 November 2024).

[6] He, Q.J. (2024) 'Research on Factors Influencing Medical Insurance Cost in the US', Transactions on Economics Business and Management Research, 10, pp. 31-36. doi:10.62051/yj92kn72.

[7] Kwapisz, A. (2022) 'Health insurance coverage and sources of advice in entrepreneurship: Gender differences', Journal of Business Venturing Insights, 14, pp. 1-12. Available at: https://doi.org/10.1016/j.jbvi.2020.e00177 (Accessed: 13 November 2024).

[8] Li, K.T. (2021) 'Regional Disparity of Medical Insurance Fund Expenditure——Based on Dagum Gini Method', Public Finance Research Journal, 2, pp. 62-71. Available at: https://cstj.cqvip.com/Qikan/Article/Detail?id=7105015056 (Accessed: 10 November 2024).

[9] Norris, K.C. (2016) 'Health Insurance and Blood Pressure Control', Journal of the American Heart Association, 5(12), pp. 1-5. doi:10.1161/jaha.116.005130.

[10] Orji, U., Ukwandu, E. (2023) Machine learning for an explainable cost prediction of medical insurance, pp. 1-42. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2311.14139&site=eds-live&scope=site (Accessed: 12 November 2024).

[11] Plemons, C. (2024) Individual and Family——What is a Health Insurance Claim? Available at: https://www.ehealthinsurance.com/resources/individual-and-family/what-is-a-health-insurance-claim (Accessed: 20 November 2024).

[12] Ramachandran, V., Kavitha, A.R. and Pandimeena, R. (2023) 'An Accurate Prediction of Medical Insurance Cost Using Forest Regression Algorithms', 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 21-23 December, pp. 1-4. Available at: https://ieeexplore.ieee.org/abstract/document/10452541/citations#citations (Accessed: 20 November 2024)

[13] Singhal, N., Goyal, S. and Singhal, T. (2022) 'The relationship between insurance and economic growth in Asian countries: a regional perspective', Macroeconomics and Finance in Emerging Market Economies, 15(3), pp. 301-322. Available at: https://doi.org/10.1080/17520843.2021.1957599 (Accessed: 15 November 2024).

[14] World Health Organization (2021) Global expenditure on health: public spending on the rise? Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/9789240041219 (Accessed: 10 November 2024).

[15] Xia, X. et al. (2024) 'Association of body mass index with risk of cardiometabolic disease, multimorbidity and mortality: a multi-state analysis based on the Kailuan cohort', Endocrine: International Journal of Basic and Clinical Endocrinology, 84, pp. 355–364. Available at: https://doi.org/10.1007/s12020-023-03570-w (Accessed: 15 November 2024).

[16] Yun, Y.H. et al. (2005) 'Cigarette smoking and cancer incidence risk in adult men: National Health Insurance Corporation Study', Cancer Detection and Prevention, 29(1), pp.15–24. Available at: https://doi.org/10.1016/j.cdp.2004.08.006 (Accessed: 13 November 2024).

[17] Zhang, P.F. (2019) 'Study on the Impact of Environmental Pollution on Medical Insurance Expenditure and Its Mechanism', Modern Economic Research, 10, pp. 28-37. doi: 10.13891/j.cnki.mer.2019.10.005.

[18] Zhou, L.Y. and Zhao, L.L. (2022) 'The Study on the Influence of WeChat-sports Health Factors on Health Insurance Premium Ratio', Science Technology and Industry, 22(7), pp. 253-258. Available at: https://cstj.cqvip.com/Qikan/Article/Detail?id=7107609776&from=Qikan_Search_Index (Accessed: 10 November 2024).