Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT (JISTM)
www.jistm.com

PREDICTING LIFE INSURANCE OWNERSHIP: THE ROLE OF SOCIOECONOMIC FACTORS

Ain'nul Mardhiah Dg. Md Nasir[1], Siti Nurasyikin Shamsuddin[2,*], Noriszura Ismail[3]

[1] Universiti Teknologi MARA Cawangan Negeri Sembilan Kampus Seremban, Seremban 70300, Malaysia
Email: 2023126389@student.uitm.edu.my
[2] Universiti Teknologi MARA Cawangan Negeri Sembilan Kampus Seremban, Seremban 70300, Malaysia
Email: syikin65@uitm.edu.my
[3] Department of Mathematical Sciences, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
Email: ni@ukm.edu.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 14.01.2025
Revised date: 23.01.2025
Accepted date: 27.02.2025
Published date: 20.03.2025

To cite this document:

Md. Nasir, A. M. D., Shamsuddin, S. N., & Ismail, N. (2025). Predicting Life Insurance Ownership: The Role of Socioeconomic Factors. *Journal of Information System and Technology Management, 10* (38), 252-270.

DOI: 10.35631/JISTM.1038017

Abstract:

Life insurance ownership is an important part of financial stability, yet ownership rates differ dramatically across income levels and demographic groupings. In order to encourage financial sustainability among a variety of demographics, it is crucial to comprehend the elements that influence life insurance ownership. This study utilizes logistic regression to predict the likelihood of life insurance ownership, with a particular focus on the impact of income and other socioeconomic factors. The findings show a substantial, positive association between income and the chance of carrying life insurance, with higher-income persons being much more likely to have policies. Education and marital status were also found to influence the probability of life insurance ownership. These results shed light on the socioeconomic variables influencing the purchase of life insurance by indicating that income differences are the main obstacle to life insurance accessibility. The study findings provide policymakers and insurers with recommendations for expanding life insurance coverage, especially among lower-income households. By identifying income disparities as a key barrier to accessibility, this research underscores the need for targeted strategies such as subsidized premium schemes, flexible payment options, and microinsurance products designed for affordability.

Keywords:

Insurance Ownership, Financial Sustainability, Logistic Regression, Socioeconomic

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Introduction

Life insurance considers the body and life as insured subjects (Zhang and Zhang, 2017). Therefore, life insurance is defined as a contractual arrangement between an individual, known as the policyholder, and the insurance company to protect the insured subjects. In this context of life insurance, the policyholder pays premium payments to the insurance company in exchange for maintaining the policy. The amount of money the policyholder pays for protection is called the premium. According to Mitra (2017), life insurance works by having the insurer take premium payments from the policyholder, invest them in less hazardous endeavours, and then return the premiums to the policyholder at the moment of death or maturity.

Life insurance is an important part of financial planning since it provides people and families with safety and stability in the face of unanticipated events. Despite its importance, life insurance uptake in Malaysia remains low, with ownership patterns ranging by demographic and socioeconomic status. Identifying the primary determinants of life insurance ownership is crucial to closing these gaps and increasing financial literacy.

This study uses logistic regression to model life insurance ownership decisions in Malaysia. Using a binary dependent variable that indicates the presence or absence of a life insurance policy, the study investigates the impact of household size, ethnicity, gender, residential strata, citizenship status, marital status, age group, income category, educational level, employment status, and expenditure patterns. This research uses a logit model to give insights into the probability of owning life insurance depending on individual and household factors.

The Malaysian government divides families into three income categories: the bottom 40% (B40), the middle 40% (M40), and the top 20% (T20) (Shamsuddin et al., 2023). According to studies, those in the higher-income group have more life insurance than the lower-income group, while the latter remains significantly underserved (Hao, 2023; Xin et al., 2024). Lower-income households encounter several challenges in acquiring life insurance, including limited financial resources, conflicting priorities, and a lack of understanding. Addressing these discrepancies is crucial to increasing overall financial inclusion and social safety nets in the country.

While previous studies have extensively examined the determinants of life insurance ownership, limited research has focused on the interplay between income categories and other socioeconomic factors in the Malaysian context. Hence, the primary objective of this research is to predict life insurance ownership in Malaysia by identifying demographic and socioeconomic factors that influence policy uptake. This method not only reveals the drivers of ownership but also lays the groundwork for policymakers and insurers to develop targeted measures to increase coverage among under-represented populations. The findings are expected to provide actionable insights for increasing life insurance penetration, particularly among the B40 and M40 groups. The findings are also likely to add to the current body of research on life insurance uptake and influence attempts to close the protection gap, delivering greater financial stability for Malaysians of all backgrounds.

Literature Review

This section will further discuss on previous research with respect to life insurance demand and income categories in Malaysia.

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Importance of Life Insurance

Life insurance is widely recognized as an essential financial tool that provides security against unexpected risks and income loss. According to Outreville (2014), life insurance contributes to economic stability by offering financial protection and promoting long-term savings. For individuals, it mitigates the financial burden caused by untimely deaths, particularly in low-income households where such events can push families into poverty.

In Malaysia, however, life insurance penetration remains relatively low compared to developed countries. The Malaysian Insurance Institute (MII) has reported that as of recent years, the penetration rate of life insurance and takaful (Islamic insurance) policies hovers around 50%, significantly below the government's target of 75%. This highlights the need to explore the factors influencing life insurance ownership, particularly among different income groups.

Socioeconomic Factors Influencing Life Insurance Ownership

Studies have consistently shown that socioeconomic factors such as income, education, and employment play significant roles in determining life insurance ownership. Browne and Kim (1993) highlighted that individuals with higher incomes and educational levels are more likely to purchase life insurance due to increased awareness and affordability. Similarly, employment stability provides financial certainty, encouraging people to invest in insurance policies (Beck and Webb, 2003).

In the Malaysian context, income disparity is a key issue affecting life insurance ownership. The Malaysian government categorizes income into three groups: B40 (Bottom 40%), M40 (Middle 40%), and T20 (Top 20%) (Malaysia Ministry of Economy, 2023), which serve as benchmarks for socioeconomic analysis. Households categorized as B40 often face affordability constraints, while the T20 group dominates life insurance ownership. A study by Bank Negara Malaysia (2019) revealed that the M40 group exhibits moderate participation in life insurance markets, suggesting the need for targeted strategies to cater to this middle-income segment. Demographic factors such as gender, ethnicity, age, and marital status also influence life insurance ownership. For instance, older individuals are more likely to prioritize life insurance due to increased health risks, while married individuals may see life insurance as a means to secure their family's financial future (Hammond et al., 1967). Research indicates that life insurance penetration is highest among the T20 group due to their higher disposable income and greater financial literacy. In contrast, B40 households often lack the resources to afford life insurance premiums, even for basic coverage.

Efforts to improve life insurance uptake among the B40 group have included the introduction of microinsurance and micro-Takaful products. However, uptake remains low due to limited awareness and competing financial priorities. Meanwhile, the M40 group represents a critical yet underserved segment, requiring customized product offerings that align with their income and lifestyle.

The application of logistic regression has been spread throughout many research areas. Recently, Rahmawati and Hsieh (2024) have conducted research that assesses the impact of Indonesia's national health insurance program on the utilization of maternal health services, employing logistic regression to analyze socioeconomic determinants. Apart from that, Lee et al. (2023) utilized multinomial logistic regression to examine how health insurance coverage and socioeconomic factors influence women's choices regarding birth attendants and delivery

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017
locations in Indonesia. To sum up, logistic regression was an important technique used in various fields.

Analytical Framework for Life Insurance Ownership

Researchers have employed various statistical models to study life insurance ownership, with logistic regression being one of the most widely used methods. Logit models are particularly effective for binary outcomes, such as determining whether an individual owns a life insurance policy (Maddala, 1983). Other studies have applied two-part models (TPM) to account for both the decision to purchase life insurance and the amount spent on premiums (Deb and Trivedi, 2002). Logistic regression was used to predict the probability of incurring a cost at a given period (Lao et al.,2022; Zhou et al.,2023). It was widely used in many fields, including medical and social science. In the medical profession, logistic regression was used to estimate the likelihood of individuals incurring healthcare costs (Couturier et al., 2022; Machinko et al., 2022; Chen and Liu, 2022).

The logistic regression model has also been used in data mining applications recently, including consumer preference models in retail and credit risk models in the banking sector (Jin et al., 2015). Finally, the insurance sector employed the logistic model to forecast claims made under travel insurance policies (Hamzah et al., 2021). Based on the study conducted by Giri and Chatterjee (2021), assessing the likelihood that uninsured households would obtain life insurance or that insured households would discontinue their insurance coverage was constructed through logistic regression.

Research Method

This section will discuss the research method of predicting life insurance ownership in Malaysia using a logistic regression.

This study method included three phases that began with acquiring and preparing data. Department of Statistics Malaysia (DOSM) provided data on Malaysian life insurance ownership for this study. Then, in the second phase, descriptive analysis and logistic regression were performed to model the likelihood of life insurance ownership. Following this, the performance evaluation was based on the results obtained in the second phase. The final phase of this study involved the performance evaluation and analysis to examine the significant relationship between the independent and dependent variables. The p-value in the chi-square test determined the significance. The descriptive analysis was carried out using SPSS Software version 27, while the logistic regression model was carried out using R Software. The selection of these two software is due to the reason that SPSS is often used for descriptive analysis because of its ease of use and robust GUI, while R software is preferred for logistic regression because of its flexibility, advanced statistical capabilities, and cost-effectiveness.

Phase 1: Data Acquisition and Preparation

This study begins with the acquisition of data from DOSM. The data was taken from the Household Income Survey and Household Expenditure Survey conducted in 2022. Once the data is acquired, it will go through a procedure that includes handling missing values and fixing discrepancies. This stage also determines which dependent and independent variables were employed in the study. Based on the availability of the data, the dependent variable in this study was life insurance ownership. Gender, household size, ethnicity, residential strata, citizenship status, marital status, age group, income category, educational level, employment status, and

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

expenditure patterns were independent variables of this study. Finally, the most important component of this step was to develop a binary indicator that can distinguish between zero and non-zero life insurance ownership. Table 1 describes the variables that were used in this study.

Table 1: Description of Variables

Attributes	Measurement	Category	Description
Ownership	Binary	1: Yes 0: No	Life Insurance Ownership 1: The household has purchased an insurance policy 0: The household has not purchased an insurance policy
HH_saiz	Interval	-	Household Size: Number of family members in the household
Gender	Binary	1: Male 0: Female	1: Male 0: Female
Ethnic	Nominal	1: Bumiputera 2: Chinese 3: Indian 4: Others	1: The ethnicity of the head of households is Bumiputera 2: The ethnicity of the head of household is Chinese 3: The ethnicity of the head of household is Indian 4: Ethnicity of the head of households is Others
Agegroup	Binary	1: 25-54 0: Others	1: Household age is in active worker years. 0: Household age is not in active worker years.
marital	Binary	1: Married 0: Others	1: Household marital status is Married 0: Household marital status is Others
strata	Binary	1: Urban 0: Rural	1: Residing in an urban area 0: Residing in a rural area
Citizenship	Binary	1: Malaysian 0: Non-Malaysian	1: Household citizenship status is Malaysian 0: Household citizenship status is non-Malaysian
incomecat	Nominal	1: B40 2: M40 3: T20	1: Household income is below RM 4850 (B40). 2: Household income in between RM 4850 to RM 10959 (M40). 3: Household income is above RM 10960 (T20).
Educational_ Level	Nominal	1: Primary 2: Secondary 3: Tertiary 4: No Certificate	1: Household's highest level of education is primary level. 2: Household's Highest level of education is secondary level

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

			3: Household's Highest level of education is tertiary level 4: No certificate
employment_ Status	Binary	1: Government 0: Others	1: Household is government employees (Government employees, Government pensioners) 0: Household is not a government employee
Premium	Interval	-	Monthly amount of life insurance premium purchased
expenditurelog	Interval	-	Log-transformed mean monthly household consumption expenditure
premiumlog	Interval	-	Log-transformed premium

Phase 2: Data Analysis

This section will determine the data analysis employed to examine the dataset.

Descriptive Analysis

As the initial step of the analysis, the descriptive analysis of the data was based on demographic information using a frequency table.

Binary Logistic Regression Model

Before running the logistic regression, the assumption of logistic regression is conducted to ensure the model provides reliable predictions. In measuring life insurance purchase decisions, the logistic regression assumed the dependent variable, life insurance ownership, was binary, as the outcome was either yes or no. Therefore, binomial logistic regression was used. The outcome was decided by categorising households who purchased life insurance as yes, while households who did not were categorised as no.

Furthermore, the relationship between the continuous independent variables and logit must be linear. The logit was the logarithm of the odds ratio, where p was the probability of an event occurs, as in this study is the probability of a life insurance purchase. A component-plus-residual plot (CR plot), often referred to as a partial-residual plot, is used to verify the linear connection. The horizontal axis of each panel displays the raw predictor values. The residuals and the unique contribution of continuous independent variables are plotted on the vertical axis after all other model predictors have been brought back into the component. The component's linear regression with the residual versus independent variables is shown by the dashed line. The dashed line would be a horizontal line at 0 if, after controlling for all other model predictors, there was no linear relationship between the dependent and continuous independent variables. A smoother that loosens the linearity assumption is the solid line. If the solid line in each panel precisely overlaps the dashed line, the linearity assumption of that predictor is fully satisfied (Harrel, 2015; Nahhas, 2024). In this example, age is linear, but there seems to be a slight non-linearity for waist circumference.

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Next, the assumption of logistic regression is by checking the multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated. It results in inaccurately predicting the life insurance purchase decision. The assumptions of multicollinearity were assessed for both numerical and categorical data. Therefore, multicollinearity assumptions were checked among gender, ethnicity, residential strata, age group, marital status, citizenship, income category, educational level, employment status, household size, life insurance ownership and log-transformed expenditure. Calculating the correlation matrix for all independent variables and identifying which pair of independent variables were highly correlated. Multicollinearity was assessed through Variance of Inflation Factor (VIF) and tolerance. Multicollinearity exists if VIF is greater than 10 (Shrestha, 2020). Thus, the formula for VIF shown below:

$$VIF = \frac{1}{1 - R^2} \quad (1)$$

The R^2 is the coefficient of determination for regression model which comes from the following linear regression model:

$$x_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \varepsilon \quad (2)$$

Multicollinearity exists if tolerance is below 0.2 thus, the formula to check the tolerance is shown below:

$$Tolerance = \frac{1}{VIF} \quad (3)$$

To ensure the dataset is suitable for logistic regression, it is assumed that there were no highly influential observations in the dataset. Outliers could cause trouble as they might have reduced the model's fitness in the dataset and impacted the validity and reliability of the model. Mahalanobis distance was used to check the outliers. The p-value less than 0.001 was considered as the outlier (Sağbaş & Balli, 2023).

After conducting the assumption of logistic regression, the process of this model was carried out based on the procedure described in an article from Hamzah et al. (2021). The procedure of this model is illustrated in the following step. The first step is to transform probability into odds. Therefore, the odds ratio is expressed as:

$$Odds = \frac{p}{1 - p} \quad (4)$$

p represents the probability of an event being favourable whereas for this study is the probability of life insurance ownership. Probability was equal to the number of favourable events and the total number of events. However, odds are the ratio between two probabilities. This model employs the odds ratio to measure the probability of having an event that was favourable to an outcome compared to having no event.

The second step is to determine the logit function. The natural logarithm (ln) of the odds ratio served as the foundation for the logistic regression model. If there were independent variables, the logistic regression model looked like:

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

$$logit = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \quad (5)$$

The β_0 in (5) represents the intercept while $\beta_1, \beta_2, \ldots, \beta_i$ represents the regression coefficients. $X_1, X_2, \ldots, X_i$ illustrates the value of independent variables. Equation (5) was a form known as the 'logit' function. The logit is a logarithm of odds where odds are the probability of occurrence of an event divided by the probability of non-occurrence of the event. It was imperative to note that the logistic function only outputs numbers between 0 and 1. The method most employed to estimate the probability for the logit model was the Maximum Likelihood Estimation (MLE). MLE was known when it came to estimating the β parameters. Next, applying exponential into (5) turns the odds become as follows:

$$Odds = \frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i} \quad (6)$$

Finally, the process concludes by calculating the estimation of the probability of an event favourable, life insurance ownership. Equation (7) determined the binary classification, either yes or no. Therefore, the logistic regression is represented by:

$$p = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)}} \quad (7)$$

This can also be expressed as:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)}} \quad (8)$$

After performing the procedure above, model adequacy checking was performed as it is essential to determine the reliability of logistic regression in representing the data. For this reason, an omnibus test will be conducted. The predictors were compared with no predictors added to the model to determine the outcome of the test. If the p-value was smaller than α = 0.05, it indicated that the dependent variable could be predicted more accurately using the information from the dependent variables. Variables with a p-value > 0.05 were excluded from further consideration in the multivariate model. Apart from the omnibus test, Cox and Snell R square and Nagelkerke R^2 were conducted. The model's goodness of fit is assessed using Nagelkerke R2 and Cox and Snell R square. It shows how much of the variation in that expected variable the model can account for. The Cox and Snell R square was calculated by comparing the model's log-likelihood to that of a baseline model, which ranged from 0 to less than 1. Nagelkerke R^2 was an adjusted version of Cox and Snell R square with the range from 0 to 1.

Phase 3: Model Evaluation of Logistic Regression

This section will describe the model evaluation used to determine the performance of logistic regression model.

Classification Table

Specificity (true negative) quantifies the proportion of the group that does not possess the characteristic of interest, whereas calculated sensitivity (true positive) quantifies the proportion of the group that possesses the characteristic of interest. The observed number of successes was

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017
compared to the expected number of successes, and the observed number of failures was compared to the projected number of failures. The classification table was used to evaluate model accuracy, sensitivity, specificity, and misclassification error rate.

Table 2 presents the prediction summary in matrix form. This confusion matrix evaluates the performance of the classification model. TP represents true positive where the number of actual positive (Y=1) is accurately classified while TN represents true negative illustrates number of actual negative (Y=0) classified accurately. On the contrary, FP is false positive, the number of actual negative (Y=0) is classified as positive (Y=1). FN is a false negative where the actual positive (Y=1) is falsely classified as negative (Y=0).

Table 2: Confusion Matrix

	Predicted	
Actual	**No (Y=0)**	**Yes (Y=1)**
No (Y=0)	TN	FP
Yes (Y=1)	FN	TP

Accuracy

Accuracy represents the proportion of the result correctly classified. Accuracy was expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Sensitivity

Sensitivity refers to the success event and assesses the likelihood of the right prediction if the event occurs. Sensitivity was determined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

Specificity

Specificity measures the model's ability to identify negative cases correctly (Satapathy, 2024). The formula for calculating specificity is as follows:

$$Specificity = \frac{TN}{TN + FP} \quad (11)$$

Results and Discussions

This section presented the findings and the interpretation of this study. The analysis was conducted using logistic regression to predict life insurance ownership.

Data Acquisition and Preparation

This study used IBM SPSS Statistics 27 to analyse descriptive analysis and R software to analyse logistic regression. Besides, this study used data from DOSM from the Household Expenditure Survey 2022 and the Household Income Survey 2022 to predict the life insurance purchase decisions or ownership. The first phase begins with identifying the data type and

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

determining the measurement level used. Next, the attributes have been categorised for example the original income attributes have been classified into three income categories (bottom 40%, B40, middle 40%, M40, and top 20%, T20). Besides, original age attributes have been categorized into two categories which are active workers' years and non-active workers' years. The original educational level has been classified into four categories which are tertiary, secondary, primary and no certificate. Employment status has also been classified into the government sector and others.

This phase also included a data preparation step involving handling missing values, cleaning, and modifying before modelling. The dataset started with a total of 17145 household data. However, only 17001 data have been considered for further analysis as the 144 data contain outliers and non-positive log-transformed premium. The outliers were detected by calculating the Mahalanobis distance in SPSS. The p-value or probability below 0.001 is considered an outlier and deleted from the dataset to create stable parameters (Sağbaş & Balli, 2023).

Descriptive Analysis

The section gives the results and explanation of findings using IBM SPSS statistics 27. Table 3 describes the summary of interval variable input. The average household size involved in this dataset is 3.8, with a minimum of 1 and a maximum of 11 people in households. The standard deviation of household size is 1.846, representing the variation in household size. The average age for households is 47.86, with a minimum of 15 and a maximum of 98 years old. The average monthly expenditure is RM 4,654.28, with a minimum spending of RM 615.26. The average premium amount is RM 25.74, and the maximum premium spending is RM 3,666.67.

Table 3: Interval Input Summary

Attributes	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
HH_saiz	1	11	3.8	1.846	0.632	0.186
Age	15	98	47.86	13.987	0.355	-0.439
Tot_Exp	615.2583	29684.8339	4654.2826	2928.8865	2.354	9.124
Premium	0	3666.6667	25.7771682	126.5754	9.979	153.563

Table 3 indicates that monthly expenditure and premium are positively skewed and have high kurtosis, indicating that the distribution has heavy tails or extreme outliers. Therefore, table 4 was conducted by applying a log into expenditure and premium to reduce the skewness effects and impact of outliers in the analysis. Applying log-transformed reduced the skewness of the premium from the original premium from 9.979 to 2.538. Moreover, applying log-transformed reduced the original data's kurtosis, 153.563 to 5.162.

Table 4: Log-transformed Variables Summary

Attributes	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Expenditurelog	2.78906	4.47253	3.6001321	0.23865187	0.159	0.161
Premiumlog	0	3.56427	0.2611955	0.68583570	2.538	5.162

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Table 5 below shows the summary of categorical variables. Most respondents are male, with 81.7% of the heads of households in Malaysia being male. More than half of the households are Bumiputera with 69.2%, followed by Chinese with 22.6%, Indian with 5.6% and others with 2.6%. 66% of households are between 25 to 54 years old. 72.4% of respondents are already married. Most families live in urban areas with a frequency of 12,144 (71.4%). Almost all the respondents are Malaysian, with 97.5% being Malaysian. B40 dominated the income category at 42.1%, M40 at 39.5%, and T20 at 18.4%. More than half of respondents have secondary (50.2%) as the highest educational level, followed by tertiary 31.7%, primary 10.2% and 7.9% not having any certificate. 68.4% of the respondents worked as government employees and pensioners in the government sector.

Table 5: Frequency Table for Categorical Variables

Attributes	Characteristics	Frequency	Percentage (%)
Gender	1: Male	13889	81.7
	0: Female*	3112	18.3
Ethnic	1: Bumiputera*	11763	69.2
	2: Chinese	3836	22.6
	3: Indian	960	5.6
	4: Others	442	2.6
Agegroup	1: 25-54 (Active Workers Years)	11224	66
	0: Others (non-active)*	5777	34
marital	1: Married	12316	72.4
	0: Others*	4685	27.6
strata	1: Urban	12144	71.4
	0: Rural*	4857	28.6
Citizenship	1: Malaysian	16583	97.5
	0: Non-Malaysian*	418	2.5
incomecat	1: B40*	7165	42.1
	2: M40	6714	39.5
	3: T20	3122	18.4
Educational _level	1: Primary*	1729	10.2
	2: Secondary	8536	50.2
	3: Tertiary	5397	31.7
	4: No Certificate	1339	7.9
employment _status	1: Government	11632	68.4
	0: Others*	5369	31.6
Ownership	1: Yes	2472	14.5
	0: No	14529	85.5

Note: * represents dummy variables (reference category)

Logistic Regression

This section discusses the output of logistic regression using R software.

Analysis of the Assumption

Figure 1 describes the relationship between continuous independent variables, log-transformed expenditure, household size, and life insurance ownership. Logistic regression models the

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

linear relationship between the independent variable and the natural logarithm of the odds of the outcome variables, life insurance ownership.

As shown in Figure 1, there was a linear association between household size, expenditure log, and life insurance ownership. The solid line (purple line) is a smoother one that relaxes the linearity assumption. In each panel, the solid line (purple line) falls exactly on top of the dashed line (blue dot), and then the linearity assumption is perfectly met for that predictor. Hence, it can be concluded that expenditure is linear, and household size seems to be slightly linear with the life insurance ownership.

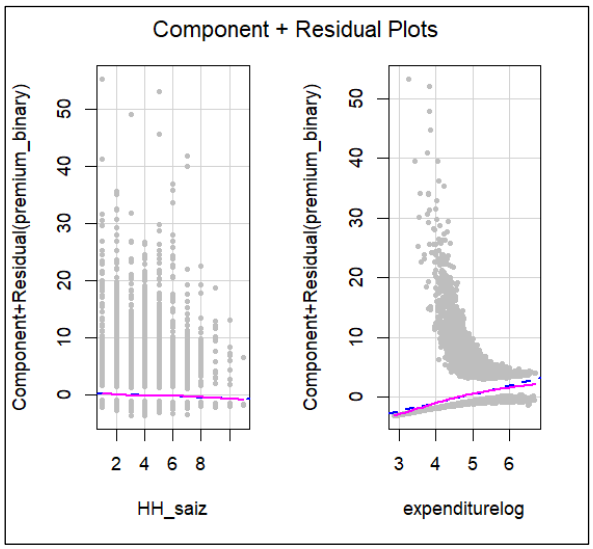

Figure 1: Residuals vs. Predictors to Check Linearity Assumption

The calculation of Mahalanobis distance in SPSS determines the outliers. Outliers were defined as data with a p-value less than 0.001 and removed from the dataset. The VIF and tolerance values are highlighted in Table 6. The tolerance levels for each variable are greater than 0.2, and the VIF values are less than 10. Multicollinearity does therefore not exist.

Table 6: Multicollinearity

Model	Tolerance	VIF
Household Size	0.756	1.322
Ethnic	0.635	1.574
Age Group	0.854	1.172
Gender	0.789	1.267
Marital Status	0.728	1.374
Strata	0.814	1.228
Citizenship	0.663	1.508
Income Category	0.362	2.761
Educational level	0.933	1.072
Employment Status	0.848	1.179
Life Insurance Ownership	0.908	1.102
Expenditurelog	0.325	3.078

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Determination of Factors Affecting Life Insurance Ownership

Table 7 shows the significant variables in logistic regression. The variable is significant if the p-value is less than 0.05. Household size, Chinese and Indian ethnicity, M40 income category, and log-transformed expenditure are highly significant as the p-value is less than 0.001. The income category of T20 and tertiary level of education is very significant as the p-value is less than 0.01. Besides, male and government employment status is significant, with a p-value of less than 0.05. The secondary level of education has weak significance since the p-value is between 0.05 and 0.1. Thus, a secondary level of education is excluded from further analysis.

Table 7: Significant Factors Affecting Life Insurance Ownership.

Attributes	Estimate (β)	Standard error	Z-value	P-value (Significant)
(Intercept)	-12.766999	0.587876	-21.717	0.000**
HH_saiz	-0.070615	0.015129	-4.668	0.000**
Ethnic2 (Chinese)	0.562998	0.054882	10.258	0.000**
Ethnic3 (Indian)	0.467542	0.092049	5.079	0.000**
Gender1 (Male)	-0.148397	0.067788	-2.189	0.02859
incomecat2 (M40)	0.408526	0.073453	5.562	0.000**
incomecat3 (T20)	0.320660	0.104167	3.078	0.00208
Educational_level2 (Secondary)	0.171129	0.103173	1.659	0.09719*
Educational_level3 (Tertiary)	0.285402	0.110630	2.580	0.00989
employment_status1 (Goverment)	0.135096	0.058443	2.312	0.02080
expenditurelog	2.894326	0.171364	16.890	0.000**

Note: * Represent the variable has weak significance, p < 0.1, ** represents variables with significant values less than 0.001

Estimation of Logistic Regression

The formula of for the Logistic Regression to predict life insurance ownership in Malaysia is given as follows:

$$\ln\left(\frac{p}{1-p}\right) = -12.766999 - 0.070615(HH\_saiz) + 0.467542(Ethnic3) - 0.148397(Gender1) + 0.408526(incomecat2) + 0.320660(incomecat3) + 0.285402(Educational\_level3) + 0.135096(employment\_status1) + 2.894326(expenditurelog) \quad (12)$$

Equation (12) shows that the estimated coefficient for household size and male gender has a negative effect. The estimated coefficient for Chinese and Indian ethnicity, M40 and T20 income category, tertiary educational level, government employment status and log-transformed expenditure has a positive effect.

Based on the results, it can be concluded that for one unit change in household size will decrease the odds of life insurance ownership by 6.82%. Chinese ethnicity will increase the odds of life insurance ownership by 75.59% compared to Bumiputera ethnicity. Indian ethnicity will increase the odds of life insurance ownership by 59.61% compared to Bumiputera ethnicity. Males will decrease the odds of life insurance ownership by 13.79% compared to Female.

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

As for income, the M40 income category will increase the odds of life insurance ownership by 50.46% compared to the B40 income category. The T20 income category will increase the odds of life insurance ownership by 37.80% compared to the B40 income category. Hence, the income category does play a significant role in determining the tendency to purchase life insurance policies.

The tertiary educational level will increase the odds of life insurance ownership by 33.03% compared to the primary educational level. Government employees have higher odds of life insurance ownership than others. For example, one unit change in the expenditure log will increase the odds of life insurance ownership by 1707.13%.

After applying the exponential in (12), thus the equation to determine the probability of a life insurance purchase is given as follows:

$$p = \frac{1}{1+e^{-\begin{pmatrix} -12.766999-0.070615(\text{HH\_saiz})+0.562998(\text{Ethnic2})+0.467542(\text{Ethnic3}) \\ -0.148397(\text{Gender1})+0.408526(\text{incomecat2})+0.320660(\text{incomecat3}) \\ +0.285402(\text{Educational\_level3})+0.135096(\text{employment\_status1}) \\ +2.894326(\text{expenditurelog}) \end{pmatrix}}} \quad (13)$$

If the p-value is greater than and equal to 0.5, the predicted value of ownership Y=1 represents the household is likely to purchase life insurance. The predicted value of ownership is Y=0 if the p-value is less than 0.5, representing the household that did not purchase life insurance.

Model Adequacy Checking

Table 8 represents the omnibus test using SPSS. The omnibus test evaluates how logistic regression explains the variation in the dependent variables. The chi-square measures how the full model fits the data compared to the null model. The chi-square statistic is 14013.285, indicating the predictors significantly have better performance due to the high chi-square statistic value. The degree of freedom is 17. This omnibus test is highly significant as the p-value is less than 0.05. Furthermore, the omnibus test also highlights the predictors that contribute to predicting the probability of life insurance ownership.

The Cox and Snell R square in Table 9 indicates the variation in that predicted variable explained by the model. Conducting Cox and Snell R square discovered 56.1% of the variance explained in the life insurance ownership. Nagelkre R^2 was an adjusted version of Cox and Snell R square, ranging from 0 to 1. The Nagelkre R^2 has a high value of 0.996, explaining the model's 99.6% of the variance in life insurance ownership. Therefore, the Nagelkre R^2 showing the model fits well.

Table 8: Omnibus Test

	Chi-Square	Degree of Freedom	Significant
Step	14013.285	17	0.000
Block	14013.285	17	0.000
Model	14013.285	17	0.000

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Table 9: Cox and Snell R square and Nagelkre R^2

Step	Cox and Snell R Square	Nagelkerke R^2
1	0.561	0.996

Model Evaluation of Logistic Regression

Classification table

Table 10 represents the classification table using a confusion matrix. The confusion matrix evaluates the performance of a binary classification model by comparing the predicted outcomes to the actual outcomes. 14414 was predicted as true negatives, indicating the model, logistic regression accurately classified the actual outcome as negatives. True negative is when the model accurately predicted the actual outcome was 0, whereas the life insurance purchase has not been made. It also predicted 2371 false positives, where the model failed to predict actual negatives (Y=0) as positives (Y=1). Besides, 121 was predicted as a false negative, whereas the actual positives were predicted as negatives. The model accurately predicted 95 as a true positive. True positive appeared when the actual outcome was 1, and life insurance was accurately predicted.

Table 10: Classification Table Summary

	Predicted	
Actual	**No (Y=0)**	**Yes (Y=1)**
No (Y=0)	14414	2371
Yes (Y=1)	121	95

Accuracy

The accuracy of the model was identified through the confusion matrix. The accuracy of the model is 85.34% correctly predicted. It shows that the model is reliable in classified instances, however, the accuracy might not fully reflect the model's ability as the dataset used is dominated by zero values to handle both positive and negative cases.

Sensitivity

The model's sensitivity indicates that roughly 43.98% of the positive cases (Y=1) in the dataset were accurately identified. However, the model struggles with the low sensitivity to detect actual positive life insurance owned due to the larger number of negative cases in the dataset compared to positive cases.

Specificity

The specificity result highlights 85.87% correctly identifying the negative cases (Y=0). This shows that the model is highly specific and provides good performance in identifying negative cases.

Summary of Model Scoring

Model scoring is performed to evaluate the accuracy and efficiency of logistic regression. Table 11 shows that the household will be predicted to purchase life insurance when the predicted value of Y=1 is more than and equal to 0.5 (Shamsuddin et al., 2023). The predicted value of Y=0 indicates that the household predicted did not purchase the life insurance. The prediction of 4 and 10 would be incorrect. Therefore, the model accuracy is 80%, and the prediction error rate is 0.2 (20%).

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Table 11: Summary of Model Scoring

Observation	Ownership (Y status)	Probability of a life insurance	Predicted for Y
1	0	0.00771	0
2	0	0.08021	0
3	1	0.58291	1
4	1	0.06300	0
5	0	0.06301	0
6	0	0.07196	0
7	1	0.53705	1
8	0	0.06668	0
9	0	0.06461	0
10	0	0.76717	1

Key Findings

Based on these findings, there are several key findings that can be highlighted. Firstly, previous research has shown that bigger families may have conflicting financial objectives, which lowers their inclination to buy life insurance (Liu and Chen, 2002; Tan et al., 2014). This is consistent with the negative association between household size and life insurance ownership. These results are supported by the projected 6.82% drop in the likelihood of owning life insurance for every unit increase in family size.

Secondly, prior research has shown that income plays a crucial role in influencing financial decisions, and this is supported by the substantial positive correlation between income and life insurance ownership. In particular, compared to the B40 (lowest 40% income group), those in the M40 and T20 income groups had higher probabilities of owning life insurance, rising by 50.46% (M40) and 37.80% (T20). Because they frequently have more money to spare, people with higher incomes might spend more in financial goods like life insurance (Tan et al., 2014; Rahman et al., 2021).

Next, in terms of education level, several studies indicate that education affects financial literacy and the capacity to comprehend and make decisions regarding insurance (Tan et al., 2014; Rahman et al., 2021; Laksono et al., 2021) support the finding that people with tertiary education have a higher likelihood of owning life insurance (33.03% higher odds compared to primary education). Greater financial knowledge and a stronger propensity to buy life insurance are frequently associated with higher educational attainment.

Apart from that, the idea that socioeconomic position and cultural characteristics play a key role in determining insurance behaviour is supported by the notable beneficial impacts of Chinese and Indian ethnic groups on life insurance ownership. For example, prior research has demonstrated that some ethnic groups are more likely to obtain life insurance and have greater levels of financial literacy (Tan et al., 2014; Deb and Norton, 2018; Zhou et al., 2023). In particular, this study found that Chinese and Indian people participate in life insurance markets at greater rates (Chinese 75.59% higher chances; Indian 59.61% higher odds).

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017

Meanwhile, research suggests that men are typically more inclined to buy insurance because of risk perceptions or financial decision-making behaviours, which is in contradiction to the lower chances of life insurance ownership for men compared to women (13.79% lower) (Chen and Liu, 2022). However, this outcome may be impacted by cultural and demographic characteristics unique to Malaysia, as well as gender roles within the households.

Furthermore, in terms of employment status, due to more consistent and reliable income, better benefits, and more financial knowledge, government employment has been associated with higher life insurance ownership (Rahman et al., 2021). The results show that government workers are more likely to have life insurance, suggesting that solid work and income are important factors when choosing a life insurance plan.

Last but not least, there may be a considerable correlation between having more money and being able to purchase life insurance, as seen by the very high probabilities of owning life insurance linked to spending (1707.13%) increase per unit change in log-transformed expenditure). Prior research has also demonstrated that people are more inclined to buy insurance products, especially life insurance, if they have greater discretionary income and spend more money (Tan et al., 2014; Rahman et al., 2021).

Conclusion and Recommendation

The purpose of this study is to predict life insurance ownership in Malaysia using logistic regression. This study discovered higher household size, Chinese and Indian ethnicity, male gender, M40 and T20 income category, tertiary educational level, government employees and higher log-transformed expenditure are the most likely to purchase life insurance. This model accurately predicts the representing dataset with 85.34%. Based to the facts, owning life insurance is greatly influenced by a number of demographic, economic, and job characteristics. The findings of this study have resulted in numerous recommendations that can be suggested for future research. Firstly, future studies should focus on the impact of handling an imbalanced dataset before identifying life insurance premium expenditure. Imbalanced datasets may influence the resulting outcome. Next, future researchers should consider other factors that may impact life insurance premium expenditure such as household health status. Furthermore, future researchers should consider behavioural and psychological variables expenditure such as trust and perceived value of life insurance in determining and providing a deeper understanding of life insurance ownership and premium expenditure.

Acknowledegment

The authors would like to thank College of Computing, Informatics and Mathematics, University Technologi MARA and Department of Statistics Malaysia (DOSM) for providing the necessary data to conduct this study. Their relevant information was crucial to the completion of this study.

References

Beck, T. & Webb, I. (2003). Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. *The World Bank Economic Review, World Bank*, vol. 17(1), pages 51-88, June.

Browne, M.J., & Kim, K. (1993). An International Analysis of Life Insurance Demand. *Journal of Risk and Insurance, 60*, 616.

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017
Chen, Y., & Liu, W. (2022). Utilization and out-of-pocket expenses of primary care among the multimorbid elderly in China: A two-part model with nationally representative data. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1057595

Couturier, V., Srivastava, S., Hidayat, B., & De Allegri, M. (2022). Out-of-Pocket expenditure and patient experience of care under-Indonesia's national health insurance: A cross-sectional facility-based study in six provinces. *The International Journal of Health Planning and Management*, *37*(S1), 79–100. https://doi.org/10.1002/hpm.3543

Deb, P., & Norton, E. C. (2018). Modeling health care expenditures and use. *Annual Review of Public Health*, *39*(1), 489–505. https://doi.org/10.1146/annurev-publhealth-040617-013517

Deb P, Trivedi PK. (2002). The structure of demand for health care: latent class versus two-part models. *J Health Econ*. Jul;21(4):601-25. 10.1016/s0167-6296(02)00008-5. PMID: 12146593.

Giri, M., & Chatterjee, D. (2021). Factors affecting changes in insured status of rural and urban households: A study over two time periods in India. *IIMB Management Review/IIMB Management Review*, *33*(4), 360–371. https://doi.org/10.1016/j.iimb.2021.12.004

Hammond, J. D., Houston, D. B., & Melander, E. R. (1967). Determinants of Household Life Insurance Premium Expenditures: An Empirical Investigation. *The Journal of Risk and Insurance*, *34*(3), 397–408. https://doi.org/10.2307/250854

Hamzah, D. A., Kalambe, A. A., Goklas, L. S., & Alkhayyat, N. G. (2021). 14.Predicting travel insurance policy claim using logistic regression. *Applied Quantitative Analysis*, *1*(1).

Hao, S. (2023). Modeling hospitalization medical expenditure of the elderly in China. *Economic Analysis and Policy*, *79*, 450–461. https://doi.org/10.1016/j.eap.2023.06.020

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer

Jin, R., Yan, F., & Zhu, J. (2015). Application of Logistic Regression Model in an Epidemiological Study. *Science Journal of Applied Mathematics and Statistics*, *3*(5), 225–229. https://doi.org/10.11648/j.sjams.20150305

Lao, C., Mondal, M., Kuper-Hommel, M., Campbell, I., & Lawrenson, R. (2022). What affects the public healthcare costs of breast cancer in New Zealand? *Asia-Pacific Journal of Clinical Oncology*, *19*(4), 482–492. https://doi.org/10.1111/ajco.13844

Lee, J. T., et al. (2023). The effect of health insurance and socioeconomic status on women's choice in birth attendant and place of delivery across regions in Indonesia: a multinomial logit analysis. *BMJ Global Health*, 8(1), e007758.

Liu, T. C., and Chen, C. S. (2002). An Analysis of Private Health Insurance Purchasing Decisions with National Health Insurance in Taiwan. *Social Science & Medicine* 55(5):755–74.

Macinko, J., Seixas, B. V., De Oliveira, C., & Lima-Costa, M. F. (2022). Private health insurance, healthcare spending and utilization among older adults: Results from the Brazilian Longitudinal Study of Aging. *The Journal of the Economics of Ageing*, *23*, 100397. https://doi.org/10.1016/j.jeoa.2022.100397

Maddala, G. (1983). Limited-Dependent and Qualitative Variables in Econometrics. *Cambridge University Press, Cambridge*. http://dx.doi.org/10.1017/CBO9780511810176

Malaysia Ministry of Economy. (2023). *RMKe-12: Rancangan Malaysia Kedua Belas* [Twelfth Malaysia Plan]. Economic Planning Unit. https://rmke12.ekonomi.gov.my/ksp/storage/fileUpload/2023/09/2023091145_main_document_ksp_rmke_12.pdf

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 10 Issue 38 (March 2025) PP. 252-270
DOI: 10.35631/JISTM.1038017
Mitra, A. (2017). Influencers of life insurance investments: Empirical evidence from Europe. *Australasian Accounting, Business and Finance Journal*, *11*(3), 87–102. https://doi.org/10.14453/aabfj.v11i3.7

Nahhas, R.W. (2024). Introduction to Regression Methods for Public Health Using R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003263197

Outreville, J.F. (2014). Risk Aversion, Risk Behavior, and Demand for Insurance: A Survey. *Journal of Insurance Issues, Western Risk and Insurance Association*, vol. 37(2), pages 158-186.

Rahman, M., Isa, C.R., Masud, M.M. *et al.* (2021). The role of financial behaviour, financial literacy, and financial stress in explaining the financial well-being of B40 group in Malaysia. *Futur Bus J* **7**, 52. https://doi.org/10.1186/s43093-021-00099-0

Rahmawati, T., & Hsieh, H. M. (2024). "Appraisal of universal health insurance and maternal health services utilization: pre- and post-context of the Jaminan Kesehatan Nasional implementation in Indonesia." *Frontiers in Public Health*, 12, 1301421.

Sağbaş, E. A., & Balli, S. (2023). Human Activity Recognition with Smartwatch Data by using Mahalanobis Distance-Based Outlier Detection and Ensemble Learning Methods. *Academic Platform Journal of Engineering and Smart Systems*, 11(3), 95–106.

Satapathy, S.K., Agrawal, P., Shah, N., Rajput, N.S. (2024). Comparative Analysis of Machine Learning and Deep Learning Algorithms for Automatic Sleep Staging Using EEG Signals. In: Pant, M., Deep, K., Nagar, A. (eds) Proceedings of the 12th International Conference on Soft Computing for Problem Solving. SocProS 2023. Lecture Notes in Networks and Systems, vol 994. Springer, Singapore. https://doi.org/10.1007/978-981-97-3180-0_16

Shamsuddin, S. N., Ismail, N., & Nur-Firyal, R. (2023). Life Insurance Prediction and Its Sustainability Using Machine Learning Approach. *Sustainability*, *15*(13), 10737. https://doi.org/10.3390/su151310737

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42.

Tan, A. K. G., Yen, S. T., Hasan, A. R., & Muhamed, K. (2014). Demand for Life Insurance in Malaysia: An Ethnic Comparison Using Household Expenditure Survey Data. *Asia-Pacific Journal of Risk and Insurance*, *8*(2). https://doi.org/10.1515/apjri-2013-0007

Xin, Y., Zhu, J., Huang, Q., Chen, Y., Chen, C., & Lu, W. (2024). Medical expenses of patients with severe mental disorders in Beijing, China. *Public Health*, *229*, 50–56. https://doi.org/10.1016/j.puhe.2024.01.022

Zhang, Q., & Zhang, X. (2017). Analysis of Influencing Factors of Life Insurance in Beijing Area. *Proceedings of the 2017 2nd International Conference on Education, Management Science and Economics (ICEMSE 2017)*. https://doi.org/10.2991/icemse-17.2017.1

Zhou, J., Wu, R., Williams, C., Emberson, J., Reith, C., Keech, A., Robson, J., Wilkinson, K., Armitage, J., Gray, A., Simes, J., Baigent, C., & Mihaylova, B. (2023). Prediction Models for Individual-Level Healthcare Costs Associated with Cardiovascular Events in the UK. *PharmacoEconomics*, *41*(5), 547–559. https://doi.org/10.1007/s40273-022-01219-6