SciMedCentral Annals of Gerontology and Geriatric Research

Research Article

Exploring the Underlying Mechanisms of Ischemic Heart Disease Post-Stroke: Insights from a 2-Year Follow-Up Investigation

Karamo Bah[1], Adama Ns Bah[2], Amadou Wurry Jallow[3], Musa Touray[4]*

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan
[2]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan
[3]Department of Medical Laboratory Science and Biotechnology, Taipei Medical University, Taipei 11031, Taiwan
[4]The University of The Gambia, School of Medicine and Allied Health Sciences, MDI Road, Kanifing P.O. Box 3530, Serrekunda, The Gambia

***Corresponding author**
Dr Musa Touray, Senior Lecturer School of Medicine and Allied Health Sciences, University of the Gambia, West Africa

Submitted: 28 November 2023
Accepted: 29 December 2023
Published: 30 December 2023

ISSN: 2378-9409

Keywords
- Ischemic heart disease
- Intensive care unit
- Ischemic stroke
- Machine learning
- Predictive modeling

Abstract

Background and Aim: Ischemic heart disease, marked by decreased blood flow to the heart muscle, is a global health concern with notable morbidity and mortality rates. Investigating novel strategies like machine learning for risk assessment, specifically among intensive care unit (ICU) admitted ischemic stroke patients, holds the potential for better outcomes. This study develops a machine learning framework to predict the 2-year risk of ischemic heart disease in ICU-admitted ischemic stroke patients, enhancing long-term prognostic accuracy.

Methods: Our study encompassed a cohort of 2,068 ischemic stroke patients admitted to the ICU from the period 2001 to 2012.We applied both a holdout strategy and a 10-fold cross-validation method during model development. Stepwise logistic regression was used to select predictors. We adopted two machine learning models such as random forest and XGBoost model for our prediction.

Results: Among the 2,068 patients, 446 had IHD during a 2-year ICU follow-up, while 1,622 did not. Baseline findings revealed that the majority of IHD patients were male (64%), with a median age of 72 years. Both XGBoost and random forest models exhibited the same discriminative power, boasting an AUROC of 93%. Notably, the top five variables in our model were platelet count, potassium levels, age, troponin T, and magnesium levels.

Conclusions: The comparative analysis highlights the superior performance of the random forest model in terms of sensitivity, specificity and accuracy, underlining its potential clinical utility for identifying high-risk patients and guiding interventions to mitigate IHD risk.

INTRODUCTION

Ischemic heart disease (IHD), also commonly referred to as coronary artery disease (CAD) or coronary heart disease (CHD), is a cardiovascular condition characterized by reduced blood flow to the heart muscle due to the narrowing or blockage of coronary arteries. These arteries supply the heart with oxygen and nutrients necessary for its proper functioning [1]. The underlying cause of IHD is atherosclerosis, a gradual buildup of fatty deposits or plaques within the coronary arteries. These plaques can restrict blood flow, leading to a reduced oxygen supply to the heart muscle. This insufficiency can result in chest pain, known as angina pectoris, or potentially cause a heart attack (myocardial infarction) when a coronary artery is completely blocked [2]. Ischemic heart disease is a leading cause of morbidity and mortality worldwide [3], responsible for a significant number of heart attacks and other cardiovascular events [4]. Common risk factors for IHD include high blood pressure, high cholesterol levels, smoking, diabetes, obesity, and a sedentary lifestyle. Diagnosis and management typically involve lifestyle modifications, medication, and, in some cases, surgical interventions like angioplasty or coronary artery bypass grafting.

Over 70% of at-risk individuals exhibit multiple IHD risk factors, while only 2%-7% of the general populace remains devoid of these risks [6]. With escalating obesity, diabetes, and metabolic syndrome prevalence, coupled with population ageing, the upward trajectory of IHD incidence is projected to persist [6]. Notably, the past two decades witnessed a steep rise in global ageing demographics [7]. The United Nations anticipates a rise in the population aged over 65 years from one in 11 in 2019 to one in six by 2050 [8]. Contemporary challenges, encompassing social connections, psychological strain, and insufficient sleep, contribute to IHD within the present generation [9]. Based on National Health and Nutrition Examination Survey (NHANES) data from 2003 to 2006, around 17.6 million Americans aged 20 or above grappled with IHD, yielding a prevalence of 7.9 per cent (9.1 per cent in men and 7 percent in women). IHD constitutes over half of all cardiovascular incidents in individuals under 75. After the age of 40, the risk of developing coronary heart disease is 49% for men and 32% for women [10].

Despite endeavors to uncover novel IHD risk factors, established ones retain a significant role [11]. Notably, these encompass dyslipidemia (elevated low-density lipoprotein cholesterol, decreased high-density lipoprotein, and elevated fasting triglycerides), hypertension, smoking, diabetes, obesity, and physical inactivity. Major prospective epidemiological studies have consistently linked these factors to heightened risk [12]. The extensive integration of Electronic Medical Record (EMR) systems presents an opportunity for an ample reservoir of data that can be employed for research purposes, including the enhanced anticipation of clinical deterioration [13]. Diverse machine learning algorithms and models have showcased their efficacy in augmenting the real-time detection of conditions like hepatitis infections, heart failure, and other illnesses [12,13,15]. This study aims to advance a machine learning framework for predicting the long-term risk of ischemic heart disease (IHD) in patients admitted to the intensive care unit (ICU) with ischemic stroke. The focus is on developing a predictive model that considers the extended 2-year follow-up period, providing a more comprehensive understanding of the evolving cardiovascular risk landscape in this patient population

METHODS

Study population

We performed a retrospective analysis utilizing the Medical Information Mart for Intensive Care (MIMIC III) V.1.4 repositories. MIMIC III is an openly accessible database containing de-identified information on 46,520 patients and 58,976 admissions at the Beth Israel Deaconess Medical Center, Boston, USA, spanning from June 1, 2001, to October 31, 2012. This dataset encompasses comprehensive details, including demographic information, admission notes, International Classification of Diseases-9th revision (ICD-9) diagnoses, laboratory test results, medication records, procedural data, fluid balance records, discharge summaries, bedside vital sign measurements, caregiver's notes, radiology reports, and survival information [16].

Our study encompassed a cohort of 2,068 ischemic stroke patients admitted to the ICU from the period 2001 to 2012.

Out of this cohort, 446 had ischemic heart disease (IHD) as dictated by established medical records and 1,622 patients were identified who did had IHD during the extended 2-year follow-up period.

Outcome Assessment

The primary focus of this study was to assess the occurrence of ischemic heart disease (IHD) as the primary outcome within a 2-year follow-up period for ICU-admitted ischemic stroke patients. The classification of patients as positive or negative cases was determined based on specific criteria:

Positive Cases: Patients were classified as positive cases if they met one or more of the following predefined criteria within the 2-year follow-up period; Presence of clinically confirmed and documented myocardial infarction events within the 2-year follow-up. Hospitalization specifically for angina pectoris within the 2-year follow-up, with documented clinical evidence of ischemic heart disease-related symptoms and confirmation by attending physicians.

Negative Cases: Patients were classified as negative cases if they did not meet any of the specific criteria outlined for positive cases within the 2-year follow-up period.

Feature extraction

To construct our model, we incorporated parameters derived from the initial admission assessments. Every parameter considered in our analysis was acquired through an exhaustive review of patient medical records archived in the MIMIC III database. The dataset encompassed structured data components pivotal for prognostication, comprising laboratory findings, demographic attributes and other viral information. Initially, we extracted 64 features from the MIMIC III structured dataset.

Feature selection

Out of the 64 features extracted, we excluded variables with missing rates exceeding 30%, which could potentially pose challenges in terms of acquisition. Subsequently, we employed insights from literature reviews to identify additional features based on their reported significance and we have 36 variables. While adopting established clinical expertise for feature selection is a prevalent approach, it may inadvertently introduce biases. Consequently, a stepwise logistic regression model was used to select features. In stepwise regression [17], a statistical technique is used to select the most relevant predictors for a binary logistic regression model. It's a systematic approach that aims to improve the model's accuracy and interpretability by automatically including or excluding predictors based on their statistical significance.

The selection criterion for including or excluding predictors was based on statistical AIC (Akaike Information Criterion) [18]. These criteria help ensure that only predictors with meaningful contributions to the model's predictive power are retained. Stepwise logistic regression can help simplify complex models, reduce over fitting, and improve model interpretability by identifying the most important predictors.

Data preprocessing

Before employing machine learning algorithms, we adopted an imputation technique [19] to handle missing values and data points beyond predefined physiological ranges. This method involves inferring missing data by replacing them with the mean or median value of a variable based on its distribution's normality. This strategy ensures the preservation of all data points, preventing information loss that might result from removing entire rows with missing values.

Post-preprocessing and ahead of model development, we leveraged the Synthetic Minority Over-sampling Technique (SMOTE) [20]. SMOTE is a specialized method used to handle missing data in datasets with imbalanced classes, unlike traditional imputation methods that focus on filling in missing values, SMOTE imputation targets the imbalance between classes by creating synthetic samples for the minority class. In imbalanced datasets, one class (the minority class) has significantly fewer samples than the other class (the majority class). This can lead to biased model outcomes. SMOTE generates synthetic samples for the minority class by interpolating existing samples. It identifies a data point from the minority class and finds its k-nearest neighbors. Then, it creates new samples by combining the features of the selected data point and its neighbors. These synthetic samples act as effective countermeasures against class imbalance. They provide the model with additional training examples for the minority class, enabling it to learn and generalize better.

Model development

This study employed the random forest model and extreme gradient boosting model (XGBoost) model. The selection of the random forest (RF) algorithm was driven by its prevalent use in clinical decision systems and its notable performance in classification tasks [21]. This algorithm employs diverse training processes on datasets to combine weak predictors into robust ones. Within the classifier employing an ensemble bagging (bootstrap aggregation) and random variable selection techniques, resulting in uncorrelated, low-bias trees [22-24].

The development of models was executed utilizing the Scikit-learn implementation (Python) [25] to yield predictions through the averaging of probability scores across the ensemble's trees, in contrast to a single-class voting approach per tree. The choice of input variables randomly selected at each split was determined as the square root of the total feature count, while the default setting was retained for the number of trees within the forest. Variable importance was gauged by evaluating the decrement in predictive performance upon omission from the model.

Chen and Guestrin [26] originally introduced the Extreme Gradient the sequential modelling of XGBoost, each decision tree builds on the outcomes of preceding iterations, culminating in a potent predictor [27]. The XGBoost algorithm assembles multiple decision trees to create the ultimate model. Notably, XGBoost substantially enhances and fine-tunes the Gradient Boosting Decision Tree method. Moreover, XGBoost consistently outperforms a standalone decision tree algorithm in terms of accuracy. Wang, Deng, and Wang [28], along with Zhao, Zheng, and Li [29], have demonstrated XGBoost's superiority over various machine learning algorithms, including support vector machine (SVM), decision trees (DT), and gradient boosting decision trees.

Of classification trees, was chosen for its merits. Each tree within the ensemble is fully grown and constructed via

Model optimization

To enhance generalization and minimize over fitting risks, we adopted a combination of holdout and cross-validation techniques. Our study dataset was divided into two segments: an 80% training set and a 20% test set. Within the training set, we conducted a 10-fold cross-validation process to facilitate feature extraction, selection, and model generation. Subsequently, we assessed model performance using the independent test set.

The model optimization, in our study, involves several key steps to ensure that our machine learning models are both effective and reliable.

Data Splitting: The first step is to divide our dataset into two main portions: a training set and a test set. We allocated 80% of our data for training and the remaining 20% for testing.

Cross-Validation: Within the training set, we applied a 10-fold cross-validation process. Cross-validation helps in utilizing training data more efficiently. Here's how it works [30]:

- The training set is divided into 10 equal parts or "folds."
- The model is trained and evaluated 10 times. In each iteration, one fold is used for validation, and the remaining nine folds are used for training.
- This process is repeated 10 times, ensuring that each fold gets a chance to be the validation set.
- The results from these 10 iterations are usually averaged to provide a more robust estimate of model performance.

The goal of this approach is to optimize our models for predictive accuracy while guarding against over fitting, where a model performs well on the training data but poorly on new data. By using both cross-validation and an independent test set, we're taking steps to ensure our models are robust and reliable when applied to real-world scenarios. This rigorous process is particularly important in healthcare applications, where model accuracy and generalization are critical.

Model evaluation

Model evaluation and validation constitute essential phases in the machine learning lifecycle, ensuring the reliability and applicability of developed algorithms. Model evaluation involves assessing a model's performance using various metrics, such as accuracy, precision, recall, F1 score, and AUC-ROC, to gauge its predictive capabilities. It enables us to understand how well the model generalizes to new and unseen data.

To evaluate the models' practical predictive capabilities within real-world scenarios, we employed various metrics, including sensitivity, specificity, accuracy, and the area under the curve (AUC). These metrics were computed using the test set, offering insights into the models' reliability and suitability for actual applications [31]. This meticulous evaluation approach ensured that our models were robust and ready for real-world deployment.

Statistical analysis

Continuous variables in each group are depicted as median (interquartile range IQR), and categorical variables are presented as absolute values and percentage n (%). Categorical data was analyzed using the Chi-square test, while continuous data was assessed using a two-sample t-test. Non-normally distributed data was expressed as median (IQR) and logarithmically transformed to approximate a normal distribution for t-tests. A significance level of $p < 0.05$ was deemed acceptable for all analyses. Statistical computations were carried out using R software version 4.3.0 (2023-04-21).

RESULTS

Basic Characteristics

Table 1 presents the baseline characteristics of the entire subject cohort. Our study comprised a total of 2,068 patients admitted to the intensive care unit (ICU) due to ischemic stroke, with an average age of 68.19 years (standard deviation 13.9), of which 1,117 were male (54%).

Table 1: Baseline characteristics of IHD and no IHD patients in the emergency department

Variables	IHD (n=446)	No IHD (n=1,622)	P value
Age (yrs.), median (IQR)	72 (64 – 78)	70 (58 – 79)	0.001*
Gender (Male) n (%)	286 (64)	831 (52)	0.001*
Platelet count (K/uL)	213 (164 – 260)	236 (183 – 304)	<0.001*
RBC (m/uL)	3.54 (3.26 – 3.91)	3.66 (3.30 – 4.11)	0.001*
WBC (K/uL)	10.05 (8.13 – 12.5)	10.85 (8.30 – 12.7)	0.795
HbA1c (%)	5.90 (5.39 – 5.90)	5.90 (5.84. – 6.00)	0.012*
Albumin (g/dL)	3.30 (3.20 – 3.50)	3.30 (2.66 – 3.47)	0.092
ALP (IU/L)	79.0 (65.7 – 86.4)	79.0 (70.7 – 92.0)	0.129
Potassium (mEq/L)	4.13 (3.93 – 4.33)	4.00 (3.81 – 4.22)	0.001*
Calcium (mg/dL)	8.60 (8.33 – 8.92)	8.60 (8.26 – 8.90)	0.066
Magnesium (mEq/L)	2.08 (1.96 – 2.23)	2.02 (1.90 – 2.16)	0.003*
Phosphate (mg/dL)	3.40 (3.09 – 3.88)	3.33 (2.92 – 3.70)	0.025*
Triglyceride (mg/dL)	119 (112 – 117)	114 (111 – 118)	0.194
LDL (mg/dL)	43.0 (41.0 – 48.3)	43 (41.0 – 44.0)	0.019*
Troponin T. (ng/dL)	0.15 (0.1 – 0.24)	0.1 (0.1 – 0.27)	0.001*
Total cholesterol (mg/dL)	160 (154 – 159)	156 (152 – 158)	0.508

IHD, Ischemic heart disease; WBC white blood cell, RBC, red blood cell; ALP, alkaline phosphate; HbA1c, Hemoglobin A1c; LDL, Low-density lipoprotein; mg/dl milligrams per deciliter; IU/L International units per litre; nanograms per deciliter (ng/dL); yrs. Years; *K/uL thousand per microliter, m/uL million per microliter, % percentage, mEq/L milliequivalents per litre; * indicates the variable is statistically significant.*

From our analysis, patients diagnosed with IHD tended to be older (median age IHD 72 years and No IHD 70 years), displaying a notable statistical difference ($P < 0.05$). Conversely, a significant difference was found between male and female patients ($P < 0.05$) (Figure 1).

Among the total ischemic stroke admissions in ICU, 446 (22%) patients experienced the development of ischemic heart disease (IHD). For those patients who experienced the development of ischemic heart disease (IHD), the median time from the onset of stroke to the occurrence of IHD was 3 days, with an interquartile range spanning from 1 to 4 days. As outlined in Table 1, patients without IHD exhibited elevated levels of platelet count, red blood cells (RBC), and white blood cells (WBC) in comparison to those with IHD. We observed significant variations across several variables, including age, platelet count, HbA1c, potassium, magnesium, phosphate, low-density lipoprotein (LDL), and troponinT ($P < 0.05$). Notably, the median values of certain laboratory parameters, such as LDL, alkaline phosphatase (ALP), albumin, and HbA1c, were consistent between both groups.

This comprehensive analysis provides insights into the contrasting characteristics and clinical markers between patients with and without IHD in the context of ischemic stroke.

Evaluation of Secondary Outcomes

In addition to the primary assessment, we investigated clinical outcomes, encompassing both the duration of hospital stay and mortality rates, within the two patient groups (Table 2).

Hyper parameters of the Models

Hyper parameters, in machine learning, are tunable settings or configurations that can be manually defined before initiating model training. These hyper parameters retain their predetermined values throughout the training process.

In our analysis, we carefully optimized select hyper parameters for each algorithm, while allowing the remaining parameters to retain their default values. Notably, for the XGBoost model, we set the learning rate at 0.1, established a maximum depth of 7, and selected 300 estimators. Conversely, in the case of the random forest model, the mtry parameter remained at its default, while we set the number of estimators to 200. Additionally, we configured the minimum samples required to split a node at 2, and the minimum samples required to form a leaf node at 1.

Table 2: Assessment of clinical outcomes

Variables	IHD	No IHD	*p-value*
Length of Stay (days) median (IQR)	3 (2 - 5)	4 (1 - 6)	0.134
Mortality n (%)	54 (12)	306 (18)	0.098

IHD, ischemic heart disease

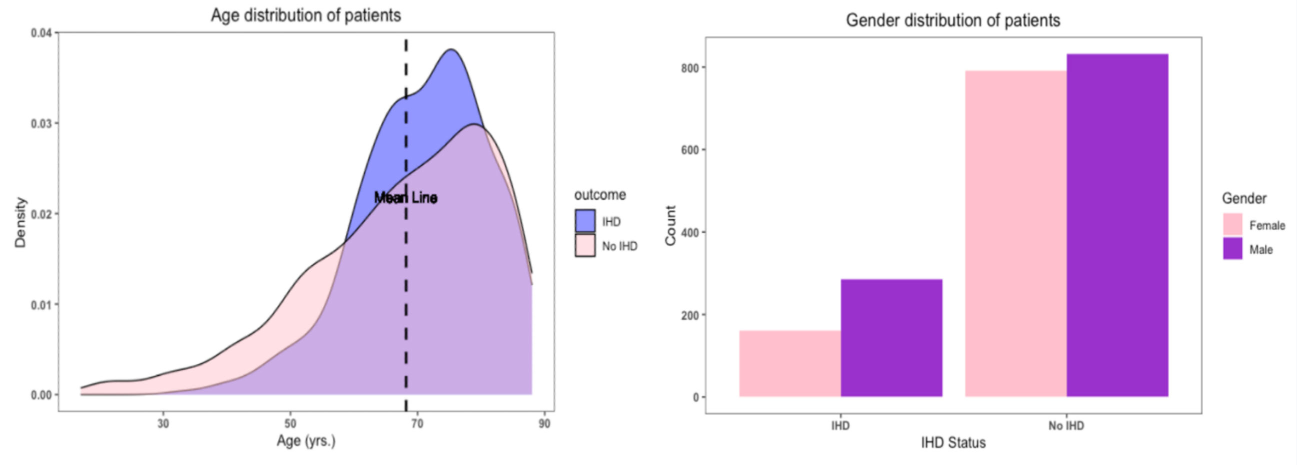

Figure 1 Age and Gender characteristics of the study participants

By strategically optimizing these hyper parameters, we aimed to strike a balance between model complexity and generalization. The chosen hyper parameters, based on careful consideration and empirical testing, significantly contribute to the models' predictive prowess while maintaining the integrity of their individual algorithms. This deliberate approach enhances the models' ability to effectively capture patterns and deliver accurate predictions.

Assessing Machine Learning Model Performance

The evaluation metrics proved instrumental in gauging the efficacy of our machine learning models. A comprehensive analysis is outlined in Table 3.

The XGBoost model demonstrated a commendable accuracy of 83%, signifying its proficiency in making correct predictions. Moreover, its sensitivity of 81% underscored the model's capacity to correctly identify positive instances, while its specificity of 86% highlighted its accuracy in recognizing negative instances. The impressive AUROC value of 93% reinforced the model's effectiveness in discerning between classes, while the F1 score of 83% emphasized a balanced trade-off between precision and recall.

In parallel, the Random Forest (RF) model exhibited remarkable performance. With an accuracy of 86%, the RF model showcased its competence in accurate classification. Its sensitivity of 84% echoed its aptitude in capturing positive instances, while its specificity of 89% highlighted a robust ability to pinpoint negative instances. The AUROC value of 93% reiterated the model's adeptness in discrimination, while the F1 score of 86% further underscored its precision-recall equilibrium.

Area Under the ROC Curve (AUROC) Analysis and Precision-Recall (PR) for the Models

The Receiver Operating Characteristic (ROC) curves provide valuable insights into the performance of our models in distinguishing between the two classes; IHD and No IHD. The AUROC, a crucial metric derived from these curves, quantifies how effectively the models separate the classes. Notably, both the XGBoost and random forest models achieved a remarkable AUROC of 93%. This robust AUROC underscores their high capacity to differentiate between IHD and No IHD cases.

Table 3: Model performance

Models	Accuracy	Sensitivity	Specificity	AUROC	PR Score	F1 Score
XGBoost	0.83	0.81	0.86	0.93	0.91	0.83
RF	0.86	0.84	0.89	0.93	0.91	0.86

XGBoost extreme gradient boosting; RF random forest; AUROC area under the receiver operating characteristic; PR, precision-recall

The compelling AUROC values obtained from these models underscore their strong discriminatory power. This is visualized in Figure 2, where the ROC curves vividly illustrate the models' ability to strike a balance between sensitivity and specificity. The convergence of these curves towards the top-left corner further accentuates the impressive discriminative performance achieved by both XGBoost and random forest.

The high AUROC values obtained across these models reinforce their efficacy in dealing with the complexity of IHD prediction. This analysis reaffirms their potential as valuable tools for identifying the presence of IHD, thereby contributing to enhanced clinical decision-making and patient care.

The Precision-Recall Curve is an essential tool for evaluating the performance of classification models, particularly in scenarios where class distribution is imbalanced. It visualizes the trade-off between precision (positive predictive value) and recall (true positive rate) as the decision threshold for class prediction varies.

Precision represents the ratio of correctly predicted positive instances to the total predicted positives. It's a valuable metric

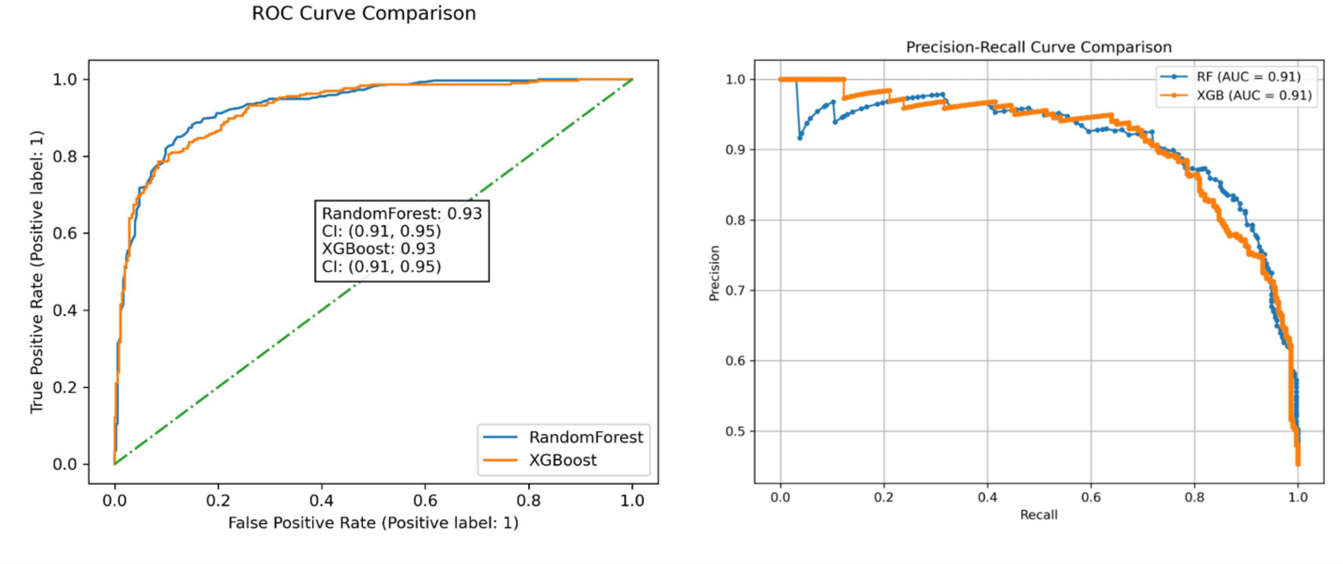

Figure 2 The receiver operating characteristic and Precision-Recall curves of the algorithms.

Table 4: Multivariate analysis of factors for IHD

Variables	Adjusted OR	95% Confidence Interval	P value
Age (yrs.)	1.02	1.01 - 1.03	<0.001*
Gender (Male) n (%)	1.70	1.22 - 2.95	<0.001*
Platelet count (K/uL)	0.99	0.99 - 0.99	<0.001*
RBC (m/uL)	0.65	0.59 - 0.85	<0.001*
WBC (K/uL)	1.03	0.97 - 1.03	0.7960
HbA1c (%)	1.16	1.03 - 1.31	0.0160*
Albumin (g/dL)	1.26	0.96 - 1.64	0.0930
ALP (IU/L)	0.99	0.99 - 1.04	0.0890
Potassium (mEq/L)	1.93	1.37 - 2.73	< 0.001*
Calcium (mg/dL)	1.26	0.98 - 1.61	0.0670
Magnesium (mEq/L)	2.09	1.27 - 3.45	0.0040*
Phosphate (mg/dL)	1.20	1.02 - 1.41	0.0260*
Triglyceride (mg/dL)	0.99	0.99 - 1.06	0.1720
LDL (mg/dL)	0.98	0.97 - 1.00	0.0170*
Troponin T. (ng/dL)	1.40	1.19 - 1.65	<0.001*
Total cholesterol (mg/dL)	1.01	0.99 - 1.05	0.5100

IHD, Ischemic heart disease; O.R, odd ratio; WBC white blood cell, RBC, red blood cell; ALP, alkaline phosphate; HbA1c, Hemoglobin A1c; LDL, Low-density lipoprotein; mg/dl milligrams per deciliter; IU/L International units per litre; nanograms per deciliter (ng/dL); yrs. Years; *K/uL thousand per microliter, m/uL million per microliter, % percentage, mEq/L milliequivalents per litre; * indicates the variable is statistically significant.*

when false positives are a concern, as it measures the model's accuracy in identifying true positives among the predicted positives.

Recall, on the other hand, represents the ratio of correctly predicted positive instances to the total actual positives. It's especially important when false negatives need to be minimized, as it measures the model's ability to capture all actual positive instances. The Precision-Recall Curve showcases the model's performance across different threshold values. A model with higher precision and recall values will exhibit a curve that approaches the top-right corner of the plot.

The area under this curve (AUC-PR) quantifies the overall performance, with a higher value indicating a better balance between precision and recall. Precision-recall curves are particularly valuable when dealing with imbalanced datasets, where negative instances significantly outweigh the positives. In such cases, accuracy alone might not provide a complete picture of the model's efficacy. By focusing on precision and recall, the Precision-Recall Curve offers a nuanced understanding of a model's ability to accurately classify positive instances while minimizing false positives and false negatives.

Variable Importance

We assessed feature importance using a random forest model. Random forest assesses feature importance by leveraging the concept of "Gini importance" or "mean decrease impurity" [22]. During the training process, each decision tree in the forest evaluates the importance of individual features by measuring how much they contribute to reducing impurity in the data.

The process involves comparing the impurity of the target variable before and after splitting a feature. Features that result in significant impurity reduction when used for splitting are considered important. The more often a feature is used for splitting across all trees in the forest and the more impurity it decreases, the higher its Gini importance score [24]. By aggregating these individual Gini importance scores across all trees in the forest, random forest generates a comprehensive assessment of feature importance. This allows practitioners to identify which features play a crucial role in making accurate predictions. This approach is particularly valuable for feature selection, aiding in model

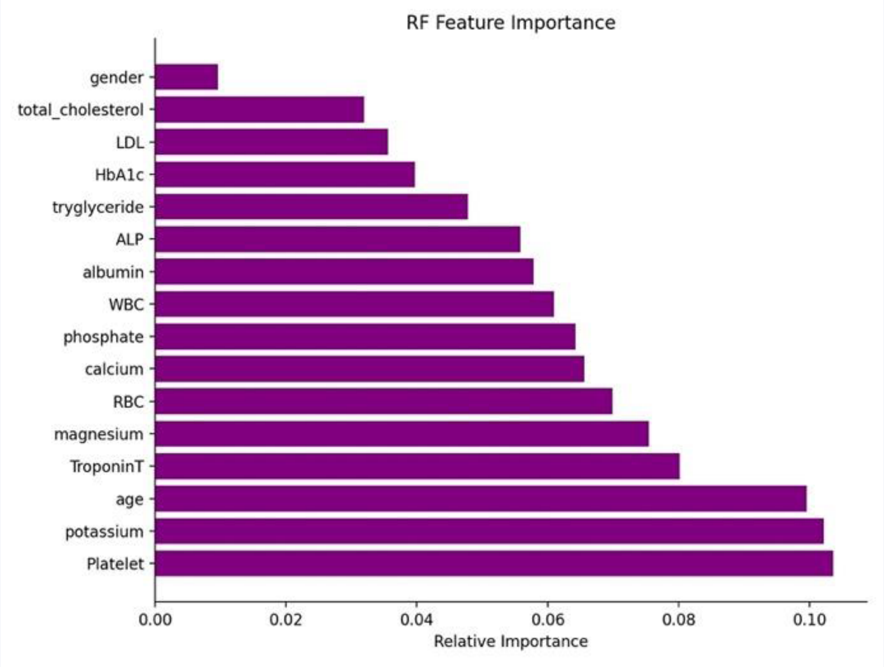

Figure 3 Feature importance contributed to the Random Forest model. WBC, white blood cell; RBC, red blood cell; ALP, alkaline phosphate; HbA1c, Hemoglobin A1c; LDL, Low density lipoprotein.

simplification and avoiding overfitting while retaining the most relevant information for prediction tasks. The top 5 features in our model are platelet count, potassium, age, troponin T and magnesium (Figure 3).

Multivariate Analysis of Factors for IHD

The results from the logistic regression multivariate analysis, presented in Table 4, shed light on the factors correlated with an elevated risk of IHD. Notably, a range of variables including age, gender, WBC, HbA1c, albumin, potassium, calcium, phosphate, troponin T, and total cholestterol demonstrated a tendency for IHD risk to increase by a factor of 1 when they augmented by one unit (OR: 1.02, 1.70, 1.03, 1.16, 1.26, 1.93, 1.26, 1.20, 1.40, 1.01) respectively. Importantly, magnesium exhibited a unique pattern with a 2-fold increase in the risk of developing IHD (OR 2.09) for each unit increase.

Conversely, platelet count, RBC, ALP, triglycerides, and LDL exhibited a significant protective effect, as their odds ratios (OR) were less than 1, indicating a reduced risk of IHD associated with these factors.

These findings underscore the nuanced interplay of various factors in IHD risk. Notably, changes in magnesium levels contribute to a significantly heightened risk. Conversely, the protective effects demonstrated by variables like platelet count, RBC, ALP, and lipid-related factors point to potentially modifiable elements that could be explored for preventive strategies in managing IHD risk.

DISCUSSION

Ischemic heart disease (IHD) poses a significant concern for patients admitted to critical care units due to ischemic stroke. The interplay between these two conditions introduces complex challenges. Ischemic stroke patients often share risk factors with IHD, such as hypertension, diabetes, and hyperlipidemia. The stress of stroke coupled with critical care interventions heightens cardiovascular strain. This underscores the need to explore the distinct risk profile and temporal dynamics of IHD in ischemic stroke patients during critical care and a subsequent 2-year monitoring phase. This investigation lays the foundation for targeted interventions to mitigate these risks and uplift patient outcomes.

Our study's key revelations are as follows: (1) The prediction system attains an impressive AUC of 0.93 in forecasting IHD risk among ischemic stroke patients using 16 clinical and demographic variables collected upon admission. (2) Platelet count, potassium, age, troponin T, and magnesium emerge as the top indicators linked to IHD risk among ICU-admitted stroke patients. (3) While both XGBoost and random forest exhibit similar AUROC curve performance, the latter excels in all other metrics. This finding suggests that, practically, the random forest model holds promise in predicting the risk of IHD among critically admitted ischemic stroke patients in the ICU. These insights illuminate a pathway for more effective clinical decisions and interventions, potentially transforming patient care in this intricate medical landscape.

Remarkably, our analysis sheds light on a set of variables that exhibit a consistent trend towards elevating the risk of stroke. This propensity is notably observed across a diverse spectrum of parameters, encompassing age, gender, WBC count, HbA1c levels, albumin levels, potassium levels, calcium levels, phosphate levels, troponin T levels, and total cholesterol levels. It is intriguing to note that, for each of these variables, an increase of one unit is associated with a parallel increase in IHD risk, denoted by odds ratios of 1.02, 1.70, 1.03, 1.16, 1.26, 1.93, 1.26, 1.20, 1.40, and 1.01, respectively.

The incremental relationship between these variables and IHD risk could offer valuable insights into the multifaceted interplay between physiological markers and the propensity for IHD. This nuanced understanding of variable impacts could significantly inform risk assessment models, thereby enabling more accurate prediction and proactive management of stroke risk in diverse clinical scenarios.

Earlier epidemiological investigations have effectively highlighted a substantial link between ischemic heart disease (IHD) and gender [32]. Earlier epidemiological investigations have effectively highlighted a substantial link between ischemic heart disease (IHD) and gender [31]. Building upon these antecedent revelations, the present study undertook a comprehensive evaluation of the interplay between gender and IHD. Notably, our findings reinforce the prevailing pattern, demonstrating that most individuals affected by IHD were male. This extends the scope of previous research and offers deeper insights into gender distribution within the IHD landscape.

Age, an incontrovertible risk factor, assumes a pivotal role in shaping susceptibility to IHD. The intricate connection between old age and IHD hinges on several factors. Firstly, cumulative exposure to traditional risk factors, including hypertension, hyperlipidemia, and diabetes, during a prolonged lifespan escalates the likelihood of developing atherosclerotic plaques, hallmark triggers of IHD. The risk of IHD increases with age, the incidence doubling with each decade after the age of 45 years and over 70% of all IHDs occur above the age of 65 [33].

Numerous preceding investigations have consistently highlighted the relationship between cholesterol levels and the risk of ischemic heart disease (IHD), portraying a gradual correlation [34]. Only a limited number of studies have delved into the comprehensive IHD risks linked to specific total cholesterol categories, particularly those encompassing values below 180 mg/dL. This gap in information might be attributed to the scarcity of individuals with TC levels below 180 mg/dL within European-origin populations [35].

In a study by Green et al., it was observed that elderly users of diuretics faced an elevated risk of stroke when their serum potassium levels were low [36]. Similarly, Smith et al. discovered a connection between low serum potassium levels and both ischemic and hemorrhagic stroke in patients receiving treatment for hypertension and IHD[37]. A recent meta-analysis dedicated to exploring the link between potassium intake and stroke risk unveiled noteworthy insights. The study revealed an inverse relationship between potassium intake and the risk of IHD [38,39].

A heightened total white blood cell (WBC) count emerges as a notable risk factor in the realm of atherosclerotic vascular disease. The involvement of WBC-derived macrophages and other phagocytes in precipitating vascular injury and fostering the advancement of atherosclerosis is widely acknowledged [40]. The evidence gleaned from numerous prospective investigations underscores a direct and autonomous correlation between WBC count and the incidence or mortality of IHD in stroke patients [41,42].

Increased levels of troponin T are observed across various acute and chronic cardiac conditions, including acute myocardial infarction (AMI), cardiac arrhythmias, and ischemic heart disease (IHD) [43]. Moreover, troponin T has demonstrated its robustness as a marker for both cardiovascular-related and overall mortality. This applies to the general population field [43] and individuals with established IHD [32].

Clinical utility

The clinical utility of our model assessing ischemic heart disease (IHD) risk in ischemic stroke patients is multifaceted and can significantly enhance patient care. Here are some of the key clinical benefits:

1. Personalized Risk Assessment - Our model offers a personalized evaluation of IHD risk for ischemic stroke patients. This tailored risk assessment empowers healthcare providers to make informed decisions that are specifically relevant to each patient's unique medical profile, optimizing interventions and treatment plans.

2. Early Intervention - By accurately predicting IHD risk, the model enables early intervention strategies. High-risk patients can be identified promptly, allowing for targeted interventions such as lifestyle modifications, medication adjustments, and close monitoring. This can potentially prevent or mitigate the development of IHD.

3. Long-Term Management - The model's risk assessment extends beyond immediate clinical decisions. It can guide long-term management strategies, helping healthcare providers and patients collaboratively plan for ongoing monitoring, risk reduction, and disease management.

4. Research and Guidelines - The insights derived from the model contribute to the body of medical knowledge. The data can be analyzed to identify trends, validate existing hypotheses, or even lead to the formulation of new research questions. Additionally, the model's outcomes could influence the development of clinical guidelines for managing IHD risk in stroke patients.

LIMITATIONS

Several noteworthy limitations warrant attention in this study. Firstly, the retrospective nature of our system's development introduces the possibility of bias. Thus, prospective validation is imperative to establish its predictive prowess. Secondly, the reliance on data solely from a single centre for both system development and evaluation underscores the necessity for additional validation with local datasets before its applicability to diverse healthcare institutions. Such expansion can bolster the robustness of our findings and instil greater confidence in the system's efficacy and generalizability. Finally, B-type natriuretic peptide (BNP) and the N-terminal fragment (NT-proBNP) are among the established biomarkers in the diagnosis of IHD, the present study did not include them because the missing values in them are high. Though including these features can increase the model performance.

CONCLUSION

Utilizing retrospective data, we have effectively crafted both a random forest and an XGBoost model by harnessing the content of MIMIC III. This innovative machine learning framework holds the potential to predict the likelihood of ischemic heart disease (IHD) development within a 2-year follow-up after ICU admission in ischemic stroke patients. Although both models exhibited comparable discriminatory power, the random forest model notably outperformed the XGBoost model across various metrics.

REFERENCES

1. Akhtar S. Ischemic heart disease. Anesthesiology Clinics of North America. 2006; 24: 461-485.
2. Moran AE, Forouzanfar MH, Roth GA, Mensah GA, Ezzati M, Murray CJL, et al. Temporal trends in ischemic heart disease mortality in 21 world regions, 1980 to 2010: the Global Burden of Disease 2010 study. Circulation. 2014; 129: 1483-1492.
3. Banerjee TK., Roy MK, Bhoi KK. Is stroke increasing in India--preventive measures that need to be implemented. J Indian Med Assoc, 2005; 103: 162, 164, 166 passim.
4. Feigin, V.L., et al., Prevention of stroke: a strategic global imperative. Nature Reviews Neurology, 2016. 12(9): p. 501-512.
5. Sampasa-Kanyinga H, Lewis RF. Frequent use of social networking sites is associated with poor psychological functioning among children and adolescents. Cyberpsychol Behav Soc Netw. 2015; 18: 380-385.
6. Desa U. World population prospects 2019: Highlights. New York (US): United Nations Department for Economic and Social Affairs. 2019; 11: 125.
7. Barquera S, Pedroza-Tobías A, Medina C, Hernández-Barrera L, Bibbins-Domingo K, Lozano R, et al. Global overview of the epidemiology of atherosclerotic cardiovascular disease. Arch Med Res. 2015; 46: 328-338.
8. Roser M, Ritchie H. Burden of disease OurWorldInData. org; 2016. 2020.
9. Virtanen M, Vahtera J, Singh-Manoux A, Elovainio M, Ferrie JE, Kivimäki M, et al. Unfavorable and favorable changes in modifiable risk factors and incidence of coronary heart disease: The Whitehall II cohort study. Int J Cardiol. 2018; 269: 7-12.
10. Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, et al. WRITING GROUP MEMBERS; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2010 update: a report from the American Heart Association. Circulation. 2010; 121: e46-e215.
11. Kuulasmaa K., Tunstall-Pedoe H, Dobson A,Fortmann S, Sans S, Tolonen H, et al. Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. Lancet. 2000; 355: 675-687.
12. McGovern PG, PankowJ S, Shahar E, Doliszny KM, Folsom AR, Blackburn H, LuepkerRV, et al. Recent trends in acute coronary heart disease—mortality, morbidity, medical care, and risk factors. N Engl J Med. 1996; 334: 884-890.
13. Das S, Eisenberg LD, House JW, Lee KL, Lusk RP, Nielsen DR, et al. Meaningful use of electronic health records in otolaryngology: recommendations from the American Academy of Otolaryngology—Head and Neck Surgery Medical Informatics Committee. Otolaryngol Head Neck Surg. 2011: 144: 135-141.
14. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of approaches for heart failure case identification from electronic health record data. JAMA cardiol. 2016; 1: 1014-1020.
15. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc. 2014; 21: 326-336.
16. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016; 3: 1-9.
17. Harrell J, Frank E, Harrell FE. Multivariable modeling strategies. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2015; 63-102.
18. Arnold TW. Uninformative parameters and model selection using Akaike's Information Criterion. The Journal of Wildlife Management. 2010; 74: 1175-1178.
19. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 2012; 28: 112-118.
20. Chawla NV, Bowyer KW, Hall LO, Philip Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16: 321-357.
21. Lakshmanaprabu S, Shankar K, Ilayaraja M, Wahid Nasir A, Vijayakumar V, Naveen Chilamkurti. Random forest for big data classification in the internet of things using optimal features. International journal of machine learning and cybernetics. 2019; 10: 2609-2618.
22. Breiman L. Random forests. Machine learning. 2001. 45: 5-32.
23. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003; 43: 1947-1958.
24. Díaz-Uriarte R, Andrés SAD. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7: 1-13.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011. 12: 2825-2830.
26. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; 785-794.

27. Zhu S, Zhu F. Cycling comfort evaluation with instrumented probe bicycle. Transportation research part A: policy and practice. 2019; 129: 217-231.

28. Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernández P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. Ann Transl Med. 2017; 5: 484.

29. Zhao T. Research on credit risk analysis based on XGBoost. Software engineering. 2018. 21: 29-32.

30. Browne MW, Cross-validation methods. J Math Psychol. 2000; 44: 108-132.

31. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8: 283-298.

32. O'Neil A, Scovelle AJ, Scovelle AJ, Kavanagh A. Gender/sex as a social determinant of cardiovascular risk. Circulation. 2018; 137: 854-864.

33. Members WG, Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, et al. Heart disease and stroke statistics—2010 update: a report from the American Heart Association. Circulation. 2010; 121: 46-215.

34. Choi JS, Song YM, Sung J. Serum total cholesterol and mortality in middle-aged Korean women. Atherosclerosis. 2007; 192: 445-447.

35. Stamler J, Wentworth D, Neaton JD. Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded?: findings in 356, 222 primary screenees of the multiple risk factor intervention trial (mrfit). JAMA. 1986; 256: 2823-2828.

36. Green D, Ropper AH, Kronmal RA, Psaty BM, Burke GL. Serum potassium level and dietary potassium intake as risk factors for stroke. Neurology. 2002; 59: 314-320.

37. Smith NL, Lemaitre RN, Heckbert SR, Kaplan RC, Tirschwell DL, Longstreth WT, et al. Serum potassium and stroke risk among treated hypertensive adults. American journal of hypertension. 2003; 16: 806-813.

38. Iso H, Stampfer MJ, Manson JE, Rexrode K, Hennekens CH, Colditz GA, et al. Prospective study of calcium, potassium, and magnesium intake and risk of stroke in women. Stroke. 1999; 30: 1772-1779.

39. Gillman MW, Cupples LA, Gagnon D, Posner BM, Ellison RC, Castelli WP, et al. Protective effect of fruits and vegetables on development of stroke in men. JAMA. 1995; 273: 1113-1117.

40. Ernst E, Hammerschmidt DE, Bagge U, Matrai A, Dormandy JA. Leukocytes and the risk of ischemic diseases. JAMA, 1987; 257: 2318-2324.

41. Berliner S, Zeltser D, Rotstein R, Fusman R, Shapira I. A leukocyte and erythrocyte adhesiveness/aggregation test to reveal the presence of smoldering inflammation and risk factors for atherosclerosis. Medical hypotheses. 2001; 57: 207-209.

42. Grimm RH, Neaton JD, Ludwig W. Prognostic importance of the white blood cell count for coronary, cancer, and all-cause mortality. JAMA. 1985; 254: 1932-1937.

43. Garg P, et al., Cardiac biomarkers of acute coronary syndrome: from history to high-sensitivity cardiac troponin. Intern Emerg Med. 2017; 12: 147-155.