

Applying Machine Learning in Cardiovascular Disease Occurrence Prediction

Po En Chung¹ , Ying Chun Li ¹ , Jong Rung Tsai² , Yu Hsin Liu ¹ , Yi Ting Chen ¹ , Pei Hung Su³

¹ National Sun Yat-sen University, Kaohsiung City, Taiwan

² Kaohsiung Municipal Ci-Jin Hospital , Kaohsiung City, Taiwan

³ Kaohsiung Gangshan Hospital , Kaohsiung City, Taiwan

P554

INTRODUCTION

- Cardiovascular disease (CVD) continues to be a leading cause of morbidity and mortality worldwide, underscoring the critical need for effective early detection and intervention strategies.
- As cardiovascular diseases are often preventable through early intervention, the establishment of robust risk assessment models is crucial.
- Although machine learning methods have been introduced to predict the occurrence of cardiovascular diseases, current research methods are mostly based on disease databases or questionnaire surveys. This presents a significant gap in the development of locally applicable health risk prediction models that can accurately identify high-risk individuals.

AIM

- This study aims to address these gaps by combining data from an adult residents' lifestyle records and health study with medical care records from a community hospital.
- By adding more risk factors and comparing the explanatory power of different machine learning analysis methods.
- We expect that using the best CVD risk prediction model in this study can not only provide early intervention and improvement measures, but also make a practical contribution to reducing the economic burden on the medical system.

METHOD

- This retrospective cohort study includes patients without clinically apparent CVD from January 2016 to December 2023, excluding participants lacking risk factor data. Patients' clinical data were linked with adult residents' lifestyle and medical care records.
- The first stage used the Cox proportional hazards model to screen for risk factors from adult residents' lifestyle records and residents' lifestyle records . Data were divided into a training dataset and a testing dataset.
- In the second stage, characteristics from the adult residents' lifestyle records medical care records were introduced into the training dataset, and four machine learning algorithms, namely logistic regression (LR), decision tree (DT), random forest (RF), and gradient boosting (GB), were used to deepen risk assessment.

RESULTS

Table1. Cox Model Measured Hazard Ratio and 95% Confidence Intervals of All-Cause Mortality of Adult Residents' Lifestyle Records

Variables	HR(95% CI) All-cause mortality	
	Age-adjusted	Multivariate-adjusted
Age		
20-40 years	1.00	1.00
41-65 years	1.40 (1.10, 1.75)**	1.35 (1.05, 1.70)**
>65 years	1.80 (1.45, 2.25)***	1.70 (1.35, 2.10)***
Gender		
Female	1.00	1.00
Male	1.65 (1.10, 2.47)**	1.45 (0.95, 2.21)
Marital status		
Married	1.00	1.00
Single	1.30(1.90, 1.80)	1.20(0.85, 1.70)
BMI (kg/m ²)		
< 18.5	1.50 (1.10, 2.05)	1.40 (1.05, 1.95)
18.5 - 24.9	1.00	1.00
>25	1.80 (1.30, 2.50)*	1.60 (1.20, 2.20)*
Smoking		
Yes	1.50 (1.10, 2.10)*	1.40 (1.05, 1.95)*
No	1.00	1.00
Alcohol		
Yes	1.25(1.10, 1.45)**	1.20(1.05, 1.35)*
No	1.00	1.00
Physical activity		
Yes	1.00	1.00
No	1.30 (1.00, 1.70)*	1.25 (0.95, 1.65)
Blood pressure (mmHg)		
SBP<120 and DBP<80	1.00	1.00
SBP≥120 and DBP≥80	2.30 (1.70, 3.10)***	2.10 (1.60, 2.85)***
Allergen		
0	1.00	1.00
1	1.20 (0.95, 1.45)	1.15 (0.90, 1.40)
2	1.30 (1.00, 1.60)	1.25 (0.95, 1.55)
≥3	1.50 (1.15, 1.90)*	1.45 (1.10, 1.85)

Table2. Cox Model Measured Hazard Ratio and 95% Confidence Intervals of All-Cause Mortality of Medical Care Records

Variables	HR(95% CI) All-cause mortality	
	Age-adjusted	Multivariate-adjusted
WBC (x10 ³ /uL)		
<3.25	1.20 (0.95, 1.50)	1.15 (0.90, 1.45)
3.25-9.16	1.00	1.00
>9.16	1.70(1.30, 2.25)***	1.65(1.25, 2.10)**
Hb		
Male < 14 / Female < 12	1.35(1.10, 1.65)**	1.30(1.05, 1.60)
Male 14-18 / Female 12-16	1.00	1.00
Male > 16 / Female > 18	1.20(0.95, 1.55)	1.15(0.90, 1.50)
MCV (fL)		
<80	1.35 (1.10, 1.65)	1.30 (1.05, 1.60)
80-100		
>100	1.50 (1.20, 1.95)**	1.45 (1.15, 1.85)**
RDW		
<12%	1.00	1.00
12-14.5%	1.40 (1.15, 1.70)	1.35 (1.10, 1.65)
>14.5%	1.60 (1.30, 2.00)	1.55 (1.25, 1.90)
Total cholesterol (mg/dL)		
<200	1.00	1.00
>200	1.40 (1.10, 1.85)**	1.35 (1.05, 1.80)**
HDL (mg/dL)		
Male > 40 / Female > 50	1.00	1.00
Male <40 / Female<50	1.40 (1.10, 1.85)**	1.35 (1.05, 1.75)**
LDL (mg/dL)		
<110	1.00	1.00
>110	1.30 (1.00, 1.65)*	1.25 (0.95, 1.60)
CRP (mg/L)		
<1	1.00	1.00
1-3	1.50 (1.20, 1.85)**	1.45 (1.15, 1.80)**
>3	1.80 (1.40, 2.25)***	1.75 (1.35, 2.15)***

Table3. Comparative of Machine Learning Algorithms for Predicting Cardiovascular Disease

Model Used	Accuracy	Recall	Precision	F1-Score
LR	0.82	0.80	0.85	0.82
DT	0.78	0.80	0.77	0.78
RF	0.84	0.85	0.87	0.86
GB	0.89	0.90	0.92	0.91

CONCLUSIONS

This study developed an effective cardiovascular disease risk prediction model using machine learning technology and integrating the data of adult residents' lifestyle records and health study with medical care records . The gradient boosting (GB) model achieved 89% accuracy , 90.0% recall, 92% precision and 91%F1-score, results which performed the best. The study results highlight the impact of questionnaire data from the adult residents' lifestyle and health study on prediction accuracy. This suggests potential improvements to cardiovascular prediction models and clinical recommendations for healthcare providers, and demonstrates the value of machine learning in disease prediction.