Original Article

A machine learning-based approach for the prediction of periprocedural myocardial infarction by using routine data

Yao Wang[1#], Kangjun Zhu[2#], Ya Li[1], Qingbo Lv[1], Guosheng Fu[1], Wenbin Zhang[1]

[1]Department of Cardiology, Key Laboratory of Biotherapy of Zhejiang Province, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China; [2]College of Computer Science and Technology, Zhejiang University of Technology, Zhejiang University, Hangzhou, China

Contributions: (I) Conception and design: Y Wang, K Zhu, G Fu, W Zhang; (II) Administrative support: G Fu, W Zhang; (III) Provision of study materials or patients: Y Li, W Zhang; (IV) Collection and assembly of data: Y Li, W Zhang; (V) Data analysis and interpretation: Y Wang, K Zhu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Wenbin Zhang; Guosheng Fu. Department of Cardiology, Key Laboratory of Biotherapy of Zhejiang Province, Sir Run Run Shaw Hospital, School of Medicine Zhejiang University, 3 East Qingchun Road, Hangzhou, China. Email: 3313011@zju.edu.cn; fugs@zju.edu.cn.

Background: Periprocedural myocardial infarction (PMI) after percutaneous coronary intervention (PCI) is associated with the bad prognosis in patients. Current approaches to predict PMI fail to identify many people who would benefit from preventive treatment, and machine learning (ML) offers opportunity to improve the performance of ML models for PMI based on the big routine data.

Methods: By using electronic medical records, we retrospectively extracted all records of patients from 2007 to 2019 in our cardiovascular center. The main enrollment criterion was that inpatients with one single coronary stenosis with stents implantation this time. The primary outcome was PMI [PMI3: cTnI >3-fold upper reference limit (URL); PMI5: cTnI >5-fold URL]. Four different ML algorithms [Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Artificial Neural Networks (ANN)] were evaluated and their diagnostic accuracy measures were compared.

Results: A total of (10,886) patients who were admitted in our hospital. PMI3 and PMI5 results were analyzed respectively. The incidence of PMI3 and PMI5 was 20.9% and 13.7%. In PMI3 Drop group, ANN (accuracy: 0.72; AUC: 0.77) showed the best power to predict the presence of PMI; In PMI3 Mean Group, RF (accuracy: 0.72; AUC: 0.77) showed the best power; In PMI5 Drop group, RF (accuracy: 0.67; AUC: 0.67) showed the best power; In PMI5 Mean group, RF (accuracy: 0.61; AUC: 0.67) showed the best power.

Conclusions: ML methods may provide accurate prediction of PMI in CAD patients, and could be used as a precise model in the preventive treatment of PMI.

Keywords: Machine learning (ML); periprocedural myocardial infarction (PMI); artificial neural networks

Submitted Jun 02, 2020. Accepted for publication Sep 28, 2020.
doi: 10.21037/cdt-20-551
View this article at: http://dx.doi.org/10.21037/cdt-20-551

Introduction

Cardiovascular diseases (CVDs) are one of the leading causes of mortality and morbidity all over the world, about 17.3 million people died in 2013 due to CVD, higher than that in 1990 (1). In the last few decades, percutaneous coronary intervention (PCI) is becoming the most popular treatment for coronary artery diseases (CADs) (2). Although technical advances in PCI process, there is still a high incident rate of periprocedural myocardial infarction (PMI) about 5–30% in the different studies (3,4). Moreover, clinical practitioners have tried their best to find the relationship between preprocedural indicators and the incidence of PMI. Unfortunately, it

isn't established that the change of a single factor or some specific factors could predict the presence of PMI.

Hopefully, machine learning (ML) provides a better way to explore the relationship between large scale baseline data and the incidence of PMI. ML can be referred as a general-purpose system with a capability of reasoning and thinking skills mimicking a human being's brain (5), and this approach relies on computers to exploit all complex and non-linear interactions across all the attributes to build the best model for the prediction of observed outcomes (6). In practice, ML has shown its power in the prediction of the risk of CVD in 10-year time period in the UK (7), prognosis in the heart failure population in the USA (8), and the length of stay in Arabian countries (9). Many investigators believe ML methods could build a more personalized and precise model based on the big dataset.

To date, these has been no large-scale investigation applying ML methods for prediction of PMI in CAD population by using preprocedural clinical data. The aim of this research is to evaluate whether ML can develop a robust prediction model for PMI. And we also plan to determine which class of ML algorithm has the highest predictive potential. We present the following study in accordance with the MDAR reporting checklist (available at http://dx.doi.org/10.21037/cdt-20-551).

Methods

Study setting and population

The dataset of this research was based on the inpatients who were admitted in department of cardiology of Sir Run Run Shaw hospital (Hangzhou, Zhejiang, China) from December 2007 to 2019 April. Since 2007, our department started to build an electronic medical system to record and evaluate the medical information for patients. Up to 2014, the electronic medical records system has been built for every department in our hospital to document outpatients and inpatients medical documentations, including demographic details, history of medical condition, bio-chemistry results, imaging impressions, primary diagnosis, prescription of drugs, records of interventions and surgeries, referrals of specialists, and following-up results. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Institutional Ethics Research Committee of Sir Run Run Shaw Hospital approved the study (approval ID: 20200224-33), and all patients provided written informed consent.

Data extraction and data preprocessing (Figure 1)

A total of 10,886 cardiac inpatients with 99 attributes were extracted from the electronic medical record system from December 2007 to 2019 April. The enrollment criteria as follows: (I) inpatients with single coronary artery stenosis (left main artery, left anterior descending artery, left circumflex artery, or right coronary artery); (II) inpatients with stent implantation this in-hospital period. The excluded criteria as follows: myocardial infarction patients or elevated pre-procedural cardiac troponin I (cTnI) or creatine kinase-MB fraction (CK-MB), PCI for more than one artery, coronary artery with thrombosis, transluminal extraction-atherectomy therapy for culprit artery, severe heart failure (EF <45% or NT-pro BNP >2,000), severe valve diseases.

About the attributes, these medical conditions related attributes were collected by experienced physicians and some laboratory results were recorded by trained technicians with standard automated machines. Also, all the results of each patient were collected in the last 24 hours before the procedure. And the cTnI and CK-MB levels were evaluated every 8 hours after the PCI, and a 24–48 hours dynamic monitoring would be acted after the procedure if necessary. However, if the data loss of an attribute was more than 10%, this attribute would be excluded. Moreover, if an attribute was belonged to another combined indicator [e.g., height and weight excluded due to body mass index (BMI)].

About the missing values, two methods were acted to do the missing data imputation, respectively (10). On the one side, the missing data would be excluded from the dataset, the remain data were used to build ML models. On the other side, mean imputation, another common approach to dealing with missing values in the ML was chose to modify the dataset.

Primary end point and definitions

The primary end point was PMI, according to the definitions of the Society for Cardiovascular Angiography and Interventions (SCAI) (11) and the Universal Definition of Myocardial Infarction (12), the Cut-Off value of cTnI for PMI was >3-fold or 5-fold upper reference limit (URL) after procedure in 48 hours. In our study, PMI3 represented the end point for cTnI >3-fold URL; PMI 5 represented the end point for cTnI >5-fold URL.

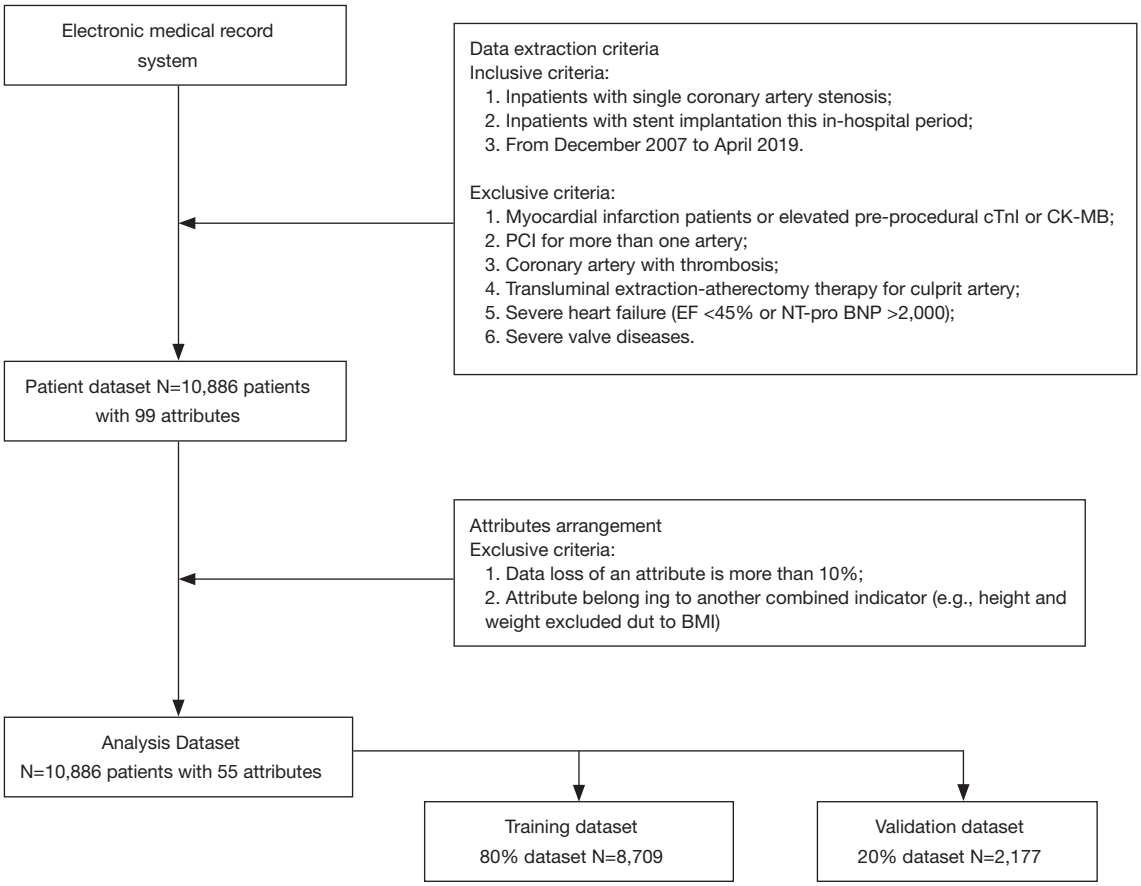

Figure 1 Patient cohort data extraction procedures.

About the attributes, the top 20 attributes were selected based on the rank determined by the information gain method for the total 55 pre-procedural variables. After choosing the final 20 attributes, ML algorithms were acted to build models to predict the PMI patients.

Feature selection and class balanced oversampling (Figure 2)

Feature selection was achieved by a method named as information gain. This method was acted in our study to rank the attributes of the dataset. Information gain measured how much information an attribute gave researchers about the outcome to be predicted. If an attribute was heterogenous in different predicted groups, it means that this attribute might be a powerful predictor in the model building process (13,14). And the value of an attribute's information gain was defined as:

$$IG(X) = H(S) - \sum_i \frac{S_i}{S} H(S_i) \quad [1]$$

Where $H(S)$ is the entropy of the dataset and $H(S_i)$ is the entropy of the i subset generated by partitioning dataset S based on a specific attribute X.

Class balanced oversampling method was another approach to balance the imbalanced dataset. Often real-word datasets were always composed of a large proportion of normal examples with a small proportion of abnormal examples (15). In our study, the population of PMI patients was much smaller than the that of negative patients, in order to have a discriminable and powerful prediction model, we used 1:1 (normal:abnormal) population to build models.

ML classification techniques

Generally, the performance of ML models can vary from

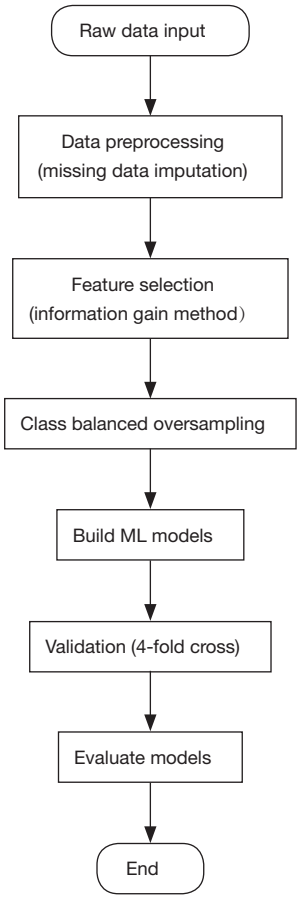

Figure 2 The flow chart of Machine learning models' building process.

different datasets, which had different variables and observed outcomes. Under this situation, four different ML algorithms [Support Vector Machine (SVM) (16), Logistic Regression (LR) (17), Random Forest (RF) (18), Artificial Neural Networks (ANN) (19)] are acted to build models in our study, moreover, the performance of each model would be determined by the area under curve (AUC) of ROC. All the algorithms were programmed and run under the Python 3.x software environment.

SVM is a classifier to aim to output an optimal hyperplane in the N-dimensional space (N is the number same as the number of attributes) to do the outcome prediction. LR is a statistical ML algorithm that classifies the data by considering outcome variables on strict final point and makes a logarithmic line that distinguishes between them. RF is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and at testing to output the class that is the mode of the classes (classification). ANN tries to mimic the human brain in order to model complicated task with many interconnected nodes just like neurons in the brain. About the model evaluation and validation, 4-fold cross validation method was acted to check the models' validation.

Statistical analysis

All data were collected by Statistical Package for the Social Science (SPSS) for macOS, version 23 (SPSS Inc., Chicago, IL, USA). Demographical characteristics of the study population were analyzed by SPSS. Categorical data were using the percentages to record, and assessed with chi-square test or Fisher's exact test. Whereas continuous data were using the means ± standard deviations to record, and assessed with Student's *t*-test. A P value less than 0.05 was considered to be statistically significant.

Results

A total of (10,886) patients who were admitted in our hospital. PMI3 and PMI5 results were analyzed respectively. The incidence of PMI3 and PMI5 was 20.9% and 13.7%. Drop imputation and Mean imputation were individually applied in the dataset to build ML models. About PMI3, all attributes in the different groups were demonstrated on the *Table 1*, as well as PMI5 in the *Table 2*. All attributes were divided into six different categories (general information, medical history, biochemistry results, blood routine examinations, medicine, and procedure factors).

As an outcome of the feature selection process, information gain method was acted to rank the attributes' importance to discriminate the non-PMI and PMI patients. *Figure 3* showed all the results in different imputation datasets. The top 5 attributes in different groups all included eGFR, VLDL-C, BMI, and LPa.

Table 3 summarized the performance of the different ML algorithms, and *Table 4* showed the performance for PMI5. Generally, the ML algorithms' performance in PMI3 was better than PMI5. Specifically, in PMI3 Drop group, ANN (accuracy: 0.72; AUC: 0.77) showed the best power to predict the presence of PMI; In PMI3 Mean Group, RF (accuracy: 0.72; AUC: 0.77) showed the best power; In PMI5 Drop group, RF (accuracy: 0.67; AUC: 0.67) showed the best power; In PMI5 Mean group, RF (accuracy: 0.61; AUC: 0.67) showed the best power.

Table 1 All attributes in PMI3 groups (PMI:non-PMI =1:1 from 10,886)

Categories	Variables	Drop			Mean		
		PMI group (N=2,274)	Non-PMI group (N=2,274)	P	PMI group (N=2,274)	Non-PMI group (N=2,274)	P
General information	Gender, male	67%	69%	0.203	69%	75%	<0.05
	Age, years	64.51±18.3	62.45±21.32	<0.05	67.85±10.05	67.71±9.88	0.311
	BMI, kg/m^2	23.44±10.81	23.98±6.11	<0.05	24.54±9.52	24.7±4.4	0.258
	SYB, mmHg	133.9±22.95	133.1±35.13	0.361	134.78±20.19	134.03±33.3	0.493
	DBP, mmHg	73.29±21.69	73.72±18.64	0.480	73.87±20.68	74.26±17.51	0.945
	UAP, yes	44%	40%	<0.05	44%	40%	<0.05
Medical history	Hyper, yes	71%	64%	<0.05	73%	69%	<0.05
	DM, yes	27%	24%	<0.05	27%	24%	<0.05
	P-CVD, yes	23%	19%	<0.05	23%	19%	<0.05
	P-PCI, yes	25%	33%	<0.05	25%	33%	<0.05
	Smoking, yes	39%	42%	<0.05	39%	42%	0.076
	Drinking, yes	24%	29%	<0.05	24%	29%	<0.05
	F-CVD, yes	13%	13%	0.756	13%	13%	0.534
Biochemistry results	TC, mmol/L	3.79±1.26	3.8±1.34	0.780	3.89±1.08	3.95±1.1	0.363
	HDL-C, mmol/L	0.99±0.31	1.01±0.34	0.141	1.02±0.27	1.05±0.28	0.223
	LDL-C, mmol/L	1.96±0.9	1.88±0.92	<0.05	2.02±0.84	1.96±0.84	<0.05
	VLDL-C, mmol/L	0.89±1.31	0.94±1.17	0.259	0.92±1.3	0.97±1.15	0.204
	TG, mmol/L	1.54±1.02	1.53±1.03	0.686	1.59±0.98	1.59±0.98	0.375
	LPa, mg/dL	25.16±26.64	21.55±24.18	<0.05	25.86±26.29	22.56±23.76	<0.05
	TB, μmol/L	12.58±6.33	13.0±6.78	<0.05	12.82±6.09	13.42±6.36	0.886
	UB, μmol/L	8.88±4.87	9.02±4.85	0.324	9.06±4.71	9.33±4.55	0.628
	CB, μmol/L	3.7±2.05	3.97±2.87	<0.05	3.77±1.98	4.09±2.78	0.597
	UA, μmol/L	349.19±120.86	339.46±115.32	<0.05	363.06±97.65	354.85±89.15	<0.05
	Cr, μmol/L	81.92±38.55	82.91±48.83	0.446	81.95±38.51	82.98±48.76	0.545
	BUN, mmol/L	5.41±2.14	5.2±1.89	<0.05	5.42±2.14	5.21±1.88	<0.05
	eGFR, mL/min	76.84±24.57	75.2±26.87	<0.05	80.2±18.53	80.85±17.06	<0.05
Blood routine examinations	WBC, ×10^9	6.57±2.05	6.41±1.68	<0.05	6.57±2.05	6.42±1.67	<0.05
	Lymphocyte, %	24.69±7.83	25.58±8.05	<0.05	24.69±7.83	25.58±8.05	0.357
	Neutrophil, %	64.03±9.83	62.15±11.32	<0.05	64.03±9.83	62.15±11.32	0.202
	Plt, ×10^9	176.82±57.13	173.45±55.71	<0.05	177.29±56.4	173.92±54.99	0.075
	MPV, fL	9.15±1.39	9.04±1.43	<0.05	9.15±1.38	9.06±1.38	0.951
	CRP, mg/L	4.36±10.5	3.06±8.99	<0.05	4.65±10.43	3.97±8.81	<0.05
	CKMB, IU	17.63±12.77	13.0±9.64	<0.05	17.63±12.77	13.0±9.64	<0.05
	FBG, mg/L	6.57±2.51	6.6±2.4	0.638	6.57±2.51	6.61±2.39	0.668

Table 1 (*continued*)

Table 1 (*continued*)

Categories	Variables	Drop			Mean		
		PMI group (N=2,274)	Non-PMI group (N=2,274)	P	PMI group (N=2,274)	Non-PMI group (N=2,274)	P
Medicine	anti-Hyper Med, yes	85%	83%	<0.05	85%	83%	<0.05
	Statins, yes	98%	98%	0.674	98%	98%	0.394
	anti-Plt Med, yes	100%	100%	<0.05	100%	100%	<0.05
	Trimetazidine, yes	25%	20%	<0.05	25%	20%	<0.05
	Fibrates, yes	0%	0%	0.466	0%	0%	0.145
	Cilostazol, yes	2%	1%	<0.05	2%	1%	<0.05
	Warfarin, yes	1%	1%	0.857	1%	1%	0.719
	PPI, yes	49%	38%	<0.05	49%	38%	<0.05
	Ezetimibe, years	8%	5%	<0.05	8%	5%	<0.05
Procedure factors	FFR, IVUS, OCT, yes	13%	8%	<0.05	13%	8%	<0.05
	CTO, yes	11%	6%	<0.05	11%	6%	<0.05
	ACC/AHA TypeB2C, yes	39%	18%	<0.05	39%	18%	<0.05
	Left coronary artery, yes	26%	30%	<0.05	26%	30%	<0.05
	Total length of stents, mm	49.66±24.92	33.16±20.15	<0.05	49.66±24.92	33.16±20.15	<0.05
	Number of stents, N	1.71±0.83	1.39±0.67	<0.05	1.71±0.83	1.39±0.67	<0.05
	Diameter of stent ≥2.5 mm	90	93	<0.05	90	93	<0.05
	Calcification, yes	16%	7%	<0.05	16%	7%	<0.05
	PCI without dilation, yes	11%	13%	0.072	11%	13%	<0.05

P value: t-statistic testing between negative group and positive group, <0.05 means significant statistically. BMI, body mass index; SYB, systolic blood pressure; DBP, diastolic blood pressure; UAP, unstable angina pectoris; Hyper, hypertension; DM, diabetes mellitus; P-CVD, past history of cerebral-or-cardiovascular diseases; P-PCI, past history of percutaneous coronary intervention; F-CVD, family history of cerebral-or-cardiovascular diseases; TC, total cholesterol; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipid cholesterol; VLDL-C, very low density lipid cholesterol; TG, triglyceride; LPa, lipid protein alpha; TB, total bilirubin; UB, unconjugated bilirubin; CB, conjugated bilirubin; UA, uric acid; Cr, creatinine; BUN, blood urea nitrogen; eGFR, estimated glomerular filtration rate; WBC, white blood cells; Plt, platelet; MPV, mean platelet volume; CRP, C-reactive protein; CKMB, Creatine Kinase MB; FBG, fibrinogen; anti-Hyper Med, anti-hypertension medicine; anti-Plt Med, anti-platelet medicine; PPI, proton-pump inhibitors; FFR, fractional flow reserve; IVUS, intravascular ultrasound; OCT, optical coherence tomography; CTO, chronic total occlusions.

Table 2 All attributes in PMI5 groups (PMI:non-PMI =1:1 from 10,886)

Categories	Variables	Drop			Mean		
		PMI group (N=1,494)	Non-PMI group (N=1,494)	P	PMI group (N=1,494)	Non-PMI group (N=1,494)	P
General information	Gender, male%	67%	68%	0.532	69%	73%	0.089
	Age, years	64.38±18.53	63.51±20.44	0.223	67.8±10.21	68.14±9.84	0.355
	BMI, kg/m^2	23.29±6.18	23.88±6.78	<0.05	24.41±3.4	24.74±4.96	<0.05
	SYB, mmHg	133.83±23.42	132.37±22.28	0.082	134.71±20.68	133.35±19.14	0.061
	DBP, mmHg	73.52±25.08	73.57±21.03	0.946	74.15±24.12	74.16±19.96	0.992
Medical history	UAP, yes	40%	44%	0.167	40%	44%	<0.05
	Hyper, yes	71%	66%	<0.05	73%	70%	<0.05
	DM, yes	27%	26%	0.619	27%	26%	0.890
	P-CVD, yes	23%	21%	0.058	23%	21%	0.058
	P-PCI, yes	25%	39%	<0.05	25%	39%	<0.05
	Smoking, yes	38%	41%	0.108	38%	41%	0.096
	Drinking, yes	24%	28%	<0.05	24%	28%	<0.05
	F-CVD, yes	13%	12%	0.349	13%	12%	0.349
Biochemistry results	TC, mmol/L	3.77±1.27	3.86±1.33	0.058	3.88±1.09	4.0±1.11	<0.05
	HDL-C, mmol/L	0.98±0.31	1.03±0.35	<0.05	1.01±0.26	1.06±0.29	<0.05
	LDL-C, mmol/L	1.96±0.9	1.92±0.92	0.278	2.02±0.84	2.0±0.84	0.540
	VLDL-C, mmol/L	0.89±1.34	0.91±1.1	0.669	0.92±1.33	0.95±1.09	0.553
	TG, mmol/L	1.54±1.04	1.58±1.11	0.369	1.59±1.01	1.64±1.07	0.226
	LPa, mg/dL	25.72±27.22	22.13±24.04	<0.05	26.47±26.83	23.07±23.63	<0.05
	TB, μmol/L	12.67±6.46	13.04±6.97	0.133	12.91±6.22	13.42±6.58	<0.05
	UB, μmol/L	8.93±4.93	9.13±4.8	0.258	9.11±4.76	9.42±4.51	0.071
	CB, μmol/L	3.74±2.12	3.9±3.23	0.112	3.81±2.05	4.01±3.15	<0.05
	UA, μmol/L	347.92±120.14	344.7±115.29	0.454	361.84±96.84	359.31±89.83	0.459
	Cr, μmol/L	82.37±44.16	84.08±55.49	0.35	82.42±44.11	84.08±55.49	0.365
	BUN, mmol/L	5.44±2.28	5.28±1.98	<0.05	5.45±2.27	5.28±1.98	<0.05
	eGFR, mL/min	76.57±25.34	75.15±25.72	0.128	80.31±18.81	80.01±17.19	0.646
Blood routine examinations	WBC, ×10^9	6.63±2.07	6.43±1.7	<0.05	6.63±2.07	6.43±1.69	<0.05
	Lymphocyte, %	24.34±7.79	25.59±8.07	<0.05	24.34±7.79	25.59±8.07	<0.05
	Neutrophil, %	64.37±9.95	62.53±10.93	<0.05	64.37±9.95	62.53±10.93	<0.05
	Plt, ×10^9	176.13±57.28	174.03±53.12	0.300	176.72±56.36	174.5±52.35	0.266
	MPV, fL	9.17±1.41	8.99±1.43	<0.05	9.17±1.41	9.01±1.37	<0.05
	CRP, mg/L	4.36±10.5	3.06±8.99	<0.05	5.01±11.93	4.22±9.21	<0.05
	CKMB, IU	18.25±14.47	13.87±8.35	<0.05	18.25±14.47	13.87±8.35	<0.05
	FBG, mg/L	6.67±2.59	6.55±2.43	0.178	6.68±2.58	6.55±2.43	0.163

Table 2 (*continued*)

Table 2 (*continued*)

Categories	Variables	Drop			Mean		
		PMI group (N=1,494)	Non-PMI group (N=1,494)	P	PMI group (N=1,494)	Non-PMI group (N=1,494)	P
Medicine	Anti-Hyper Med, yes	86%	84%	0.169	86%	84%	0.169
	Statins, yes	98%	98%	0.449	98%	98%	0.374
	Anti-Plt Med, yes	100%	100%	0.076			0.076
	Trimetazidine, yes	25%	22%	0.064	25%	22%	0.062
	Fibrates, yes	0%	0%	0.781	0%	0%	0.782
	Cilostazol, yes	2%	2%	0.229	2%	2%	0.228
	Warfarin, yes	1%	1%	1.000	1%	1%	0.999
	PPI, yes	51%	38%	<0.05	51%	38%	<0.05
	Ezetimibe, yes	8%	6%	0.051	8%	6%	0.050
Procedure factors	FFR, IVUS, OCT, yes	14%	10%	<0.05	14%	10%	<0.05
	CTO, yes	12%	5%	<0.05	12%	5%	<0.05
	ACC/AHA TypeB2C, yes	39%	23%	<0.05	39%	23%	<0.05
	Left coronary artery, yes	26%	31%	<0.05	26%	31%	<0.05
	Total length of stents, mm	51.47±25.37	33.34±20.23	<0.05	51.47±25.37	33.34±20.23	<0.05
	Number of stents, N	1.75±0.85	1.37±0.662	<0.05	1.75±0.85	1.37±0.662	<0.05
	Diameter of stent ≥2.5 mm	91	92	0.146	91	92	0.146
	Calcification, yes	17%	8%	<0.05	17%	8%	<0.05
	PCI without dilation, yes	11%	11%	0.772	11%	11%	0.772

P value: t-statistic testing between negative group and positive group, <0.05 means significant statistically. BMI, body mass index; SYB, systolic blood pressure; DBP, diastolic blood pressure; UAP, unstable angina pectoris; Hyper, hypertension; DM, diabetes mellitus; P-CVD, past history of cerebral-or-cardiovascular diseases; P-PCI, past history of percutaneous coronary intervention; F-CVD, family history of cerebral-or-cardiovascular diseases; TC, total cholesterol; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipid cholesterol; VLDL-C, very low density lipid cholesterol; TG, triglyceride; LPa, lipid protein alpha; TB, total bilirubin; UB, unconjugated bilirubin; CB, conjugated bilirubin; UA, uric acid; Cr, creatinine; BUN, blood urea nitrogen; eGFR, estimated glomerular filtration rate; WBC, white blood cells; Plt, platelet; MPV, mean platelet volume; CRP, C-reactive protein; CKMB, Creatine Kinase MB; FBG, fibrinogen; anti-Hyper Med, anti-hypertension medicine; anti-Plt Med, anti-platelet medicine; PPI, proton-pump inhibitors; FFR, fractional flow reserve; IVUS, intravascular ultrasound; OCT, optical coherence tomography; CTO, chronic total occlusions.

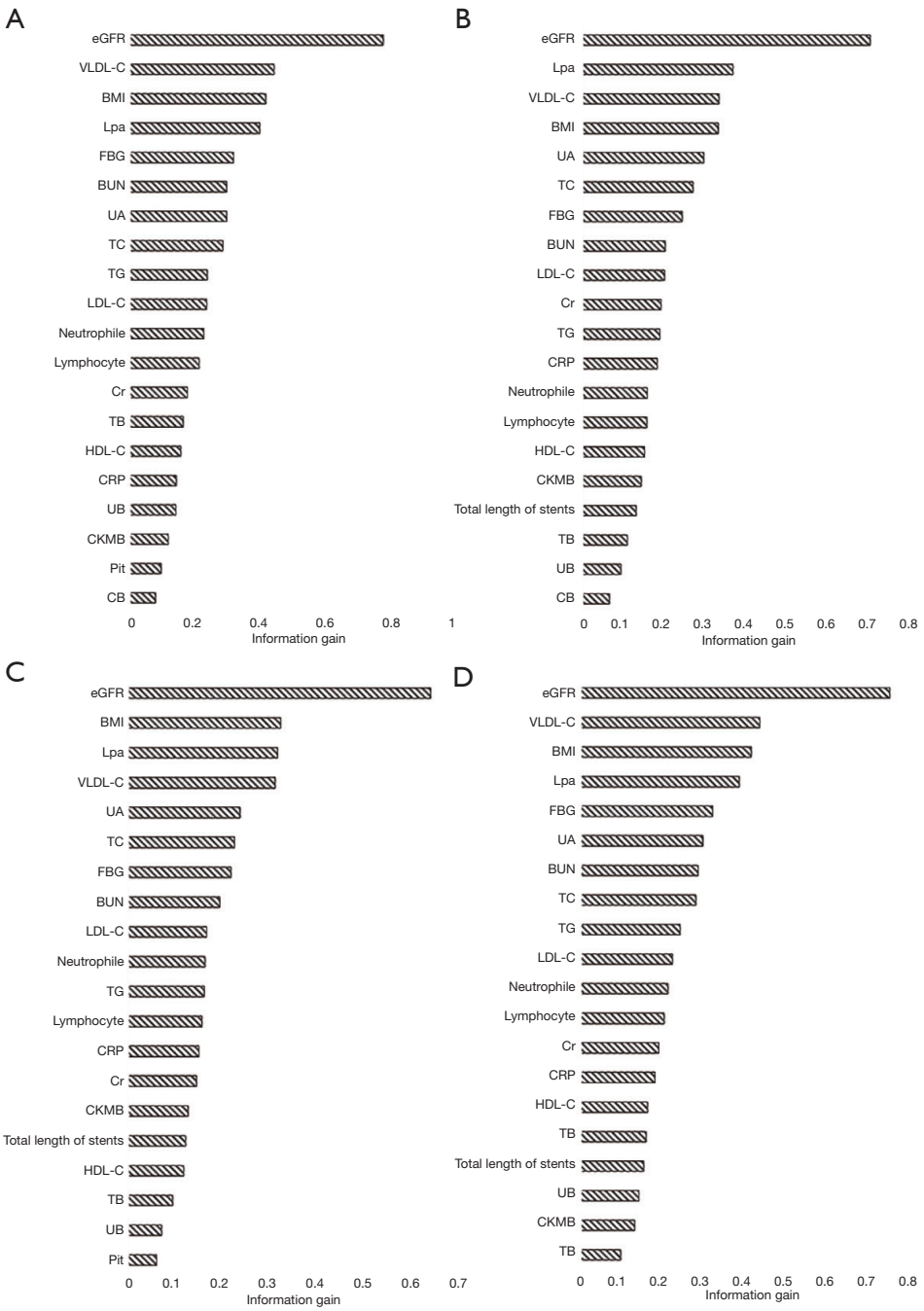

Figure 3 The information gain of top 20 attributes in different data imputation methods. (A) Drop PMI3; (B) mean PMI3; (C) drop PMI5; (D) mean PMI5. BMI, body mass index; SYB, systolic blood pressure; DBP, diastolic blood pressure; UAP, unstable angina pectoris; Hyper, hypertension; DM, diabetes mellitus; P-CVD, past history of cerebral-or-cardiovascular diseases; P-PCI, past history of percutaneous coronary intervention; F-CVD, family history of cerebral-or-cardiovascular diseases; TC, total cholesterol; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipid cholesterol; VLDL-C, very low density lipid cholesterol; TG, triglyceride; LPa, lipid protein alpha; TB, total bilirubin; UB, unconjugated bilirubin; CB, conjugated bilirubin; UA, uric acid; Cr, creatinine; BUN, blood urea nitrogen; eGFR, estimated glomerular filtration rate; WBC, white blood cells; Plt, platelet; MPV, mean platelet volume; CRP, C-reactive protein; CKMB, Creatine Kinase MB; FBG, fibrinogen; anti-Hyper Med, anti-hypertension medicine; anti-Plt Med, anti-platelet medicine; PPI, proton-pump inhibitors; FFR, fractional flow reserve; IVUS, intravascular ultrasound; OCT, optical coherence tomography; CTO, chronic total occlusions.

Table 3 The performance of different machine learning models in PMI3

Models	Drop				Mean			
	SVM	LR	RF	ANN	SVM	LR	RF	ANN
Accuracy	0.69	0.71	0.71	0.72	0.69	0.70	0.72	0.70
Sensitivity	0.73	0.67	0.70	0.72	0.73	0.66	0.70	0.70
PPV	0.67	0.72	0.71	0.71	0.67	0.71	0.72	0.72
F1-score	0.69	0.69	0.70	0.72	0.70	0.68	0.71	0.71
AUC	0.76	0.76	0.77	0.77	0.76	0.76	0.77	0.76

SVM, Support Vector Machine; LR, Logistic Regression; RF, Random Forest; ANN, artificial neural network; PPV, positive predictive value; AUC, area under curve.

Table 4 The performance of different machine learning models in PMI5

Models	Drop				Mean			
	SVM	LR	RF	ANN	SVM	LR	RF	ANN
Accuracy	0.60	0.62	0.67	0.62	0.61	0.62	0.61	0.61
Sensitivity	0.62	0.62	0.68	0.56	0.66	0.62	0.66	0.61
PPV	0.58	0.59	0.59	0.62	0.58	0.59	0.58	0.61
F1-score	0.60	0.61	0.63	0.59	0.62	0.61	0.62	0.61
AUC	0.66	0.65	0.67	0.65	0.66	0.65	0.67	0.65

SVM, Support Vector Machine; LR, Logistic Regression; RF, Random Forest; ANN, Artificial neural network; PPV, positive predictive value; AUC, area under curve.

Discussion

This study aimed to develop a ML-based model to predict the presence of PMI in CAD patients. A total of 10,886 patients were recruited in our study, six different categories data (general information, medical history, biochemistry results, blood routine examinations, medicine, and procedure factors) were used to consist of the dataset, four different algorithms (SVM, LR, RF, and ANN) were acted to build models to predict the PMI. ANN and RF showed the most accurate performance in PMI3 and PMI5.

Firstly, eGFR, VLDL-C, BMI, and LPa were four top indicators based on the analysis of information gain. In the previous studies, periprocedural TnT levels of eGFR <60 mL/min/1.73 m^2 patients who received PCI were three times higher than the normal limit (20). About VLDL-C and LPa, remnant cholesterol could be valuable and independent predictor for PMI in diabetic patients (21), and non-HDL-C showed its power for the diagnosis of myocardial injury post procedurally (22). BMI was a combined index to assess people's the degree of obesity, and an article reported that overweight patients had a better prognosis after primary angioplasty compared with other BMI groups. Information gain was a powerful method to explore the relationship between attributes and outcomes, not only just an important step to preprocess the data, but also may be used to assess the degree of relationship.

Secondly, different missing data imputation would lead to different results. In our study, two different methods for missing data imputation were acted to preprocess the data set, Drop the missing data and Median substitution. It is very hard to say which treatment for the missing data was better, moreover, our results also didn't show the evidence for which method was better. Besides two ways in our study, other methods like k-nearest neighbors imputation and frequent values imputation, were also recommended in some articles (23,24). What's more, multiple imputation was another way to replace missing data (25) in order to get a better dataset in some degree.

Thirdly, ML ANN and RF algorithms had a better power for prediction of PMI than ML SVM and LR

algorithms in this research. In particular, there was no one-size-fits-all ML model. Literally, although RF and ANN had a better performance in this study, we cannot conclude that these two models would be better in other datasets. For example, in a Korean Study, deep neural networks were the best ML to predict the risk of CVD (26). And RF algorithm also showed its better power for prediction of in-hospital length (27). ANN showed its better performance in determining the risk of CVD in the UK population (7). Moreover, some articles reported that ML models were better than the traditional risk systems to predict the special outcomes, such as mortality of heart failure (8), acute coronary syndrome requiring revascularization (28), and risk of CVD mentioned before.

Fourthly, the performance of four ML models to predict PMI3 was better than that to predict PMI5. The difference between PMI3 dataset and PMI5 dataset was the data scale. Although we had 10,886 patients in our original dataset, about 80% of data were belonging to the negative group, after the oversampling step, ML algorithms cannot build a great model based on a small-scale dataset. In other words, datasets were the fundamental essential for a better prediction rather than the methods, including ML algorithms.

Limitations

Generally, there were several limitations of this current research. Firstly, as mentioned before, the dataset was from one-single health organization instead of several different centers. Secondly, it was acknowledged that the "black box" nature of ML models could be impossible for the interpretation of ML models. Thirdly, if the data loss of an attribute reached 10%, the attribute was removed from the dataset. This preprocess step would cause some biases before we knew the specific variable was significant for the prediction or not.

Conclusions

ML methods provide accurate prediction of PMI in CAD patients, and could be used as a precise model in the preventive treatment of PMI.

Acknowledgments

Xue Lv has given us a help with language editing.
Funding: This work was supported by the National Natural Science Foundation of China (81500212 and 81800212) and Zhejiang Natural Science Foundation (LY18H020007 and LQ17H020002).

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at http://dx.doi.org/10.21037/cdt-20-551

Data Sharing Statement: Available at http://dx.doi.org/10.21037/cdt-20-551

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/cdt-20-551). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Institutional Ethics Research Committee of Sir Run Run Shaw Hospital approved the study (approval ID: 20200224-33), and all patients provided written informed consent.

References

1. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015;385:117-71.
2. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association [published correction appears

in Circulation. 2018 Mar 20;137(12):e493]. Circulation 2018;137:e67-e492.
3. Herrmann J. Peri-procedural myocardial injury: 2005 Update. Eur Heart J 2005;26:2493-519.
4. Testa L, Van Gaal WJ, Biondi Zoccai GGL, et al. Myocardial infarction after percutaneous coronary intervention: A meta-analysis of troponin elevation applying the new universal definition. QJM 2009;102:369-78.
5. Seetharam K, Shrestha S, Sengupta PP. Artificial Intelligence in Cardiovascular Medicine. Curr Treat Options Cardiovasc Med 2019;21:25.
6. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: A methodology review. J Biomed Inform 2002;35:352-9.
7. Weng SF, Reps J, Kai J, et al. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12:e0174944.
8. Panahiazar M, Taslimitehrani V, Pereira N, et al. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform 2015;216:40-4.
9. Daghistani TA, Elshawi R, Sakr S, et al. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. Int J Cardiol 2019;288:140-7.
10. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 2003;17:519-33.
11. Moussa ID, Klein LW, Shah B, et al. Consideration of a New Definition of Clinically Relevant Myocardial Infarction After Coronary Revascularization. J Am Coll Cardiol 2013;62:1563-70.
12. Thygesen K, Alpert JS, Jaffe AS, et al. Fourth universal definition of myocardial infarction (2018). Russ J Cardiol 2019;24:107-38.
13. Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manag 2006;42:155-65.
14. Kent J. Information Gain and a General Measure of Correlation. Biometrika 1983;70:163-73.
15. Mordant N, Delour J, Léveque E, Arnéodo A, et al. Long time correlations in lagrangian dynamics: a key to intermittency in turbulence. Phys Rev Lett 2002;89:254502.
16. Huang X, Maier A, Hornegger J, et al. Indefinite kernels in least squares support vector machines and principal component analysis. Appl Comput Harmon Anal 2017;43:162-72.
17. Kleinbaum DG, Klein M. Logistic Regression A Self-Learning Text Second Edition., 2002.
18. Classification and regression by randomForest. R News 2:18-22. Forest 2001;23(3).
19. Yao X, Member S. Evolving Artificial Neural Networks. Proceedings of the IEEE 1999;87:1423-47.
20. Kumagai S, Ishii H, Amano T, et al. Impact of chronic kidney disease on the incidence of peri-procedural myocardial injury in patients undergoing elective stent implantation. Nephrol Dial Transplant 2012;27:1059-63.
21. Zeng RX, Li S, Zhang MZ, et al. Remnant cholesterol predicts periprocedural myocardial injury following percutaneous coronary intervention in poorly-controlled type 2 diabetes. J Cardiol 2017;70:113-20.
22. Maadani M, Abdi S, Parchami-Ghazaee S, et al. Relationship between pre-procedural serum lipid profile and post-procedural myocardial injury in patients undergoing elective percutaneous coronary intervention. Res Cardiovasc Med 2013;2:169.
23. Courrieu P, Rey A. Missing data imputation Chapter 25. Behav Res Methods 2011;43:310-30.
24. Planned M, Protocols M, Design TPM, et al. Statistical Issues : What Happens When Data 2015;I:760-97.
25. Faris PD, Ghali WA, Brant R, et al. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidemiol 2002;55:184-91.
26. Kim J, Kang U, Lee Y. Statistics and deep belief network-based cardiovascular risk prediction. Healthc Inform Res 2017;23:169-75.
27. Timóteo AT, Ramos R, Toste A, et al. Impact of body mass index in the results after primary angioplasty in patients with ST segment elevation acute myocardial infarction. Acute Card Care 2011;13:123-8.
28. Noh YK, Park JY, Choi BG, et al. A Machine learning-based approach for the prediction of acute coronary syndrome requiring revascularization. J Med Syst 2019;43:253.

Cite this article as: Wang Y, Zhu K, Li Y, Lv Q, Fu G, Zhang W. A machine learning-based approach for the prediction of periprocedural myocardial infarction by using routine data. Cardiovasc Diagn Ther 2020;10(5):1313-1324. doi: 10.21037/cdt-20-551

