International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1231 - 1236
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0169

Improving Cardiovascular Disease Forecasting with Machine Learning and Electronic Medical Record Data Characteristics Within a Local Healthcare Network

Mrs. Sapana Bhushan Raghuwanshi[1], Dr. Nilesh Ashok Suryawanshi[2]
[1]Research Scholar, Gangamai College of Engineering Nagoan, Dhule, Maharashtra, India.
[2]Assistant Professor, Gangamai College of Engineering Nagoan, Dhule, Maharashtra, India.
Emails: *sapana.salunkhe99@gmail.com[1], nileshsuryawanshipatil088@gmail.com[2]*

Abstract
The PCE Risk Calculator, developed by the ACC/AHA, is frequently utilized in the United States for the purpose of averting the onset of Atherosclerotic cardiovascular disease (ASCVD) via first-line defense strategies. However, this calculator may not accurately estimate risk for certain populations, potentially leading to either under- or over-estimation of risk. We have created calculator for ASCVD risk specific to a population by leveraging advanced Machine Learning (ML) techniques and Electronic Medical Record (EMR) data. Our study involved comparing predictive accuracy of our calculator with PCE calculator. Between January 1, 2009, and April 30, 2020, data was gathered from 101,110 distinct EMRs of patients who were actively receiving treatment. Patient datasets underwent machine learning techniques containing Longitudinal (LT) and Cross-Sectional (CS) features, or solely CS features, derived from laboratory values and vital statistics. The models' effectiveness was assessed using fresh price metric (Screened Cases Percentage @Sensitivity level). In terms of prediction accuracy, every ML model that was tested performed better than the PCE risk calculator. Area Under Curve (AUC) score of 0.902 was obtained by Random Forest (RF) ML technique when CS and LT characteristics were combined (RF-LTC). Our machine learning model only needed to screen 43% of patients in order to identify 90% of positive ASCVD cases, in contrast to the PCE risk calculator, which required screening 69% of patients. Prediction models created using ML techniques reduce the amount number of tests necessary to forecast ASCVD and increase the accuracy of ASCVD prediction when compared to using PCE calculator alone. The combination of LT and CS features in these ML models leads to a significant enhancement in comparing the ASCVD prediction to utilizing CS features exclusively.

Keywords: *Cardiovascular disease; Electronic health record; Machine learning; Mass screening; Risk.*

1. Introduction
ASCVD, also known as atherosclerotic cardiovascular disease, has significant implications for both health and the economy on a global scale [1]. Offering ASCVD risk scores to high-risk individuals can help lower their risk by prompting preventive interventions and reducing the need for diagnostic procedures [2-4]. In the field of clinical practice, there is a growing focus on enhancing ASCVD risk scores to enhance cost efficiency and minimize the potential risks linked to costly or invasive examinations [5&6]. The present recommendations utilize pooled cohort equations (PCE) for the assessment of the 10-year risk of developing hard ASCVD [non-fatal myocardial infarction (MI), non-fatal stroke, or fatal from CHD and for informing treatment strategies [4&7]. Different models are currently utilized in clinical settings to forecast ASCVD. Nevertheless, these models have the potential to inaccurately assess risk by either underestimating or overestimating the true observed risk in populations with varying comorbidities or demographic and socioeconomic factors [2&5, 8-10]. Risk calculators that are readily accessible are now being integrated into EMR platforms with decision support [2] capabilities. Nevertheless, there is still a requirement for tools that can offer a more extensive assessment of ASCVD risk over a period of time in Figure 1.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1231 - 1236
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0169

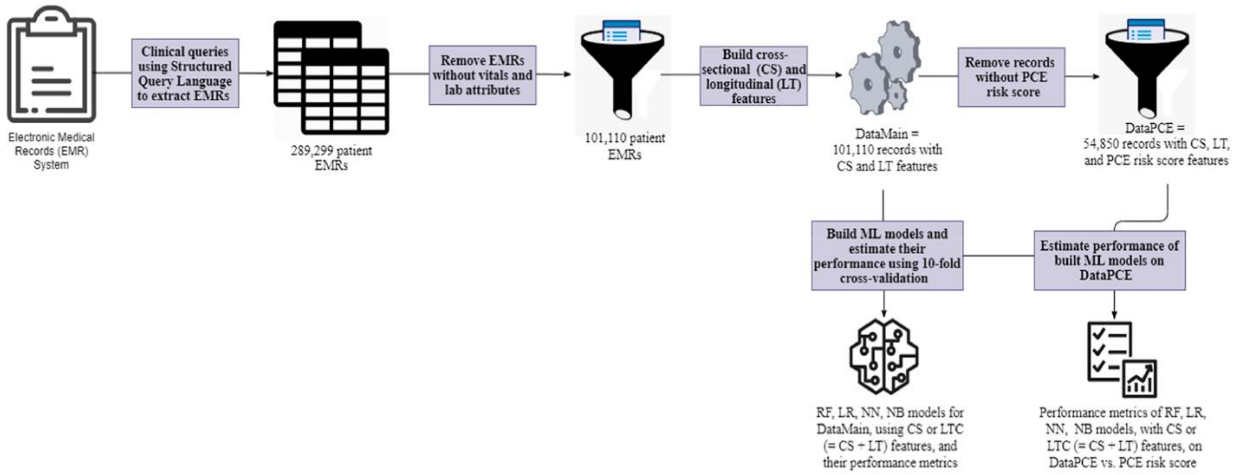

Figure 1 The Procedure for Constructing and Testing ML Models

ML has been implemented in a system for making medical decisions and has consistently demonstrated comparable or superior performance when compared to risk prediction decisions made by humans in the field of cardiology [5&8]. Longitudinal data extracted from electronic medical record (EMR) systems can aid in the utilization of machine learning (ML) techniques to enhance clinical risk assessment for atherosclerotic cardiovascular disease (ASCVD) [8]. Prior research has concentrated on the integration of ML algorithms in the detection of clinical characteristics from coronary artery calcium (CAC) scores to ASCVD-related events [8-11]. Despite the effectiveness of CAC as a reliable and economical tool for reclassifying ASCVD risk [4,5,8], obstacles continue to hinder its widespread adoption. We have developed a clinically-focused ASCVD prediction model in this study that successfully fills in the gaps in the available techniques.

2. Method

A longitudinal study was carried out retrospectively, utilizing datasets from distinct electronic medical records (EMRs) of patients who are still alive that are managed through a local healthcare network in the United States. We used structured query language (SQL) to extract data from the medical records of patients at St. Elizabeth Health Care System (Kentucky, USA) who had clinical visits between January 1, 2009, and April 30, 2020, which included the measurement of low-density lipoprotein cholesterol (LDL-C). We did this by using an interactive EMR-driven clinical decision support system. The study searched for patients with documented LDL-C levels and used a validated equation to estimate pretreatment levels for statin therapy. A total of 289,299 records were selected for ML models, including those with PCE risk scores. When a native coronary artery develops atherosclerotic heart disease or unstable angina, the PCE score is determined. The study allowed for comparison of ML models and longitudinal features affecting predictive accuracy. We also used the IJMEDI medical AI assessment checklist to cross-check and validate our discussion, result reports, and model design [12]. ASCVD refers to patients with CAD, CVS, or PAD who have their records marked with a 'time stamp' indicating the presence of the disease.

2.1. Cross-Sectional Features and Longitudinal Features

We also used the IJMEDI medical AI assessment checklist to cross-check and validate our discussion, result reports, and model design [12]. To capture time-sequential LT features like blood pressure, HbA1c, and lipid profile, we created an additional 63 LT features. The statistics were computed for patients who did not have a diagnosis of ASCVD, using all the data gathered during the study period. Prior to the ASCVD diagnosis, we exclusively took into account the readings documented in the EMR for patients diagnosed with ASCVD.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1231 - 1236
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0169

2.2. ML Models

In order to forecast the probability of a patient developing ASCVD, we constructed automated models using four different ML techniques: Neural networks (NN), random forests (RF), logistic regression (LR), and naïve Bayes (NB). In one case, the models were built using not more than CS features as predictors, and in another, a combination of LT features (LTC) and CS features was used as a predictor.

2.3. Screened Cases Percentage @ Sensitivity Level

A prediction model's main goal is to reduce the number of patients screened while increasing sensitivity. The Screened Cases Percentage@Sensitivity (SCP@Sensitivity) metric was developed to measure this. It shows the percentage of the population that must be screened to a particular sensitivity threshold.

$$\text{SCP @Sensitivity(S)} = \frac{|\text{Subpopulation of patients who must be screened to achieve the target sensitivity level of S}|}{|\text{Overall Patient Population}|}$$

where |X| represents the cardinality of set X.

This method can also be utilized in a clinical setting to evaluate the necessary resources for screening patients in order to reach a specific sensitivity level within the general population, potentially leading to a decrease in unnecessary testing. Theoretically, all ASCVD positive patients (designated as SCP@Sensitivity) could be identified by screening as few as 11.56% of patients in the DataPCE cohort (1). This percentage means that out of the 54,850 cases in DataPCE, there are 6339 ASCVD patients overall.

3. Results and Discussion

3.1.Results

As stated in Section 3.1.1, we conducted a thorough comparison of all models and provided their AUC scores. This analysis was carried out with the aim of identifying the best model and feature sets. Moving forward, in Section 3.1.2, we delved into an examination of the features utilized in the best model to uncover their significance. We compared the models with and without PCE in Section 3.1.3 in order to answer the question of whether adding PCE as a feature improves our models. Furthermore, in Section 3.1.4, our model in a comparison with the existing PCE calculator to determine if our model outperforms it. In Section 3.1.5, we conducted a statistical comparison of our model. Moreover, a new metric was implemented, which evaluates case percentages at a specific sensitivity threshold, as outlined in Section 3.1.6. In accordance with Section 3.1.7, we presented the performance of our best model based on its AUC score and probability threshold to help determine the ideal threshold.

3.1.1. Model Performance

The RF-LTC model demonstrated superior performance in ASCVD prediction compared to RF-CS, NN-LTC, LR-LTC, and NB-LTC models. It achieved the highest AUC of 0.902 (95% CI, 0.895–0.910). The AUC represents an overall measure that considers the probabilities connected to the ROC curve's paired sensitivity and specificity.

The neural network uses backpropagation with three layers and key parameters set at 0.01 and 0.9. Different algorithms are implemented using Scikit-learn, including Random Forest, Logistic Regression, and Naive Bayes.

3.1.2. Impact of Features Used to Build the

The RF-CS model revealed that the most significant factors were age, comorbidities, and aggregate risk scores, with LDL-C values trailing closely behind. In the LTC model, blood pressure, lipid levels, and HbA1c were the most predictive factors. Age was consistently highlighted as one of the key features in RF-CS as well as RFLTC models [13].

3.1.3. ML Comparison with the PCE Features and Without the PCE Features

To evaluate how the PCE score affects the performance of the model, we conducted an evaluation of the identical models using the DataPCE dataset. This evaluation encompassed both PCE features (PCE scores and PCE categorical) and models without PCE features. The NN-LTC model achieved the highest AUC score of 0.896, matching that of the model with PCE features. Additionally, the AUC score of the RF_LTC model is 0.894, showing a decrease of 0.006.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1231 - 1236
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0169

3.1.4. PCE Calculator Comparison with Machine Learning

The CE 10-year risk score, which is currently employed in clinical settings, was compared to the automated machine learning techniques. The DataPCE dataset was used for the comparison with the PCE score, since the PCE risk scores could only be obtained from this dataset. The ML models were built using the DataMain dataset. The ML models demonstrated superior ASCVD prediction compared to the PCE risk calculator, with an AUC of 0.712 (95% CI, 0.700–0.730). Within the DataPCE dataset, the LR-LTC model achieved the highest AUC of 0.880 (95% CI, 0.867–0.894).

3.1.5. Calibration Curves and Net Reclassification Index

The certainty of fresh observations from models in recognized classes was evaluated using the Brier score. Models with LTC features had better Brier scores than models with CS features. Comparing machine learning models was another use of continuous NRI, with RF-LTC models being significantly better than NNN, NN, LR, and LR models on DataMain. PCE had a Brier Score of 0.262.

3.1.6. Percentage of Screened Cases Between 50% and 90% of Sensitivities [SCP@0.9) And SCP@0.5] in Terms of Sensitivity

Given that risk-prediction techniques necessitate the inclusion of laboratory and other diagnostic examinations to validate or disprove the diagnosis, When the method correctly predicts a larger percentage of potentially positive cases (like ASCVD) while requiring additional testing for a smaller percentage of the population as a whole (SCP@Sensitivity), the efficacy of the method is increased. If a technique is modified to enhance its sensitivity, it could result in higher expenses, depending on the proportion of the population necessitating testing. In the DataPCE cohort, SCP@Sensitivity (0.9) might theoretically be as low as 10.40% in order to achieve a sensitivity of 90%. The PCE approach necessitates evaluating every patient with a risk score of 5% or higher. The NN-LTC algorithm, on the other hand, only needed to screen 43.4% of the total population. In terms of SCP@Sensitivity (0.90), all machine learning models outperformed the PCE score. SCP@Sensitivity(0.5) in the DataPCE group may be as low as 5.78% to cover half of the true positive cases in order to reach a sensitivity of 0.5. 25.6% of the population must be screened in order to use the PCE calculator while the RFLTC model requires screening 7.1%. All ML models outperformed the PCE in screening a smaller proportion of cases in order to reach a 50% sensitivity.

3.1.7. Probability Threshold for the Performance of the Datamain Model

When it comes to clinical practice, the person in charge of making decisions makes binary predictions about outcomes, classifying the data as either 1 (positive) or 0 (negative). Conversely, machine learning models offer a continuous prediction, displaying a risk score between 0 and 1. Because of this predicted risk score, an analysis that depends on a particular threshold value, 't,' is required. A projected risk score that is greater than or equal to the selected probability threshold value "t" is regarded as a positive prediction; a predicted risk score that is less than "t" is regarded as a negative prediction. Given that the RF-LTC model demonstrates the highest AUC values overall, we have presented the performance of the RFLTC model along with their respective threshold probability values. The purpose of this information is to help clinicians choose the best threshold based on different requirements. The 0.25 threshold yielded the best results for the RF_LTC model. Cut-off probability can be selected based on clinical necessity to reduce the number of high-risk patients who are incorrectly classified.

3.2. Discussion

A recent study has demonstrated that machine learning models incorporating both clinical and lifestyle factors outperform models that only consider lifestyle factors in predicting ASCVD risk. By analyzing 94 clinical variables, the study identified age, PCE risk score, HTN, and DM as crucial predictors for ASCVD. It underscores the significance of taking into account multiple variables and cumulative risk in ASCVD prediction, advocating for the annual evaluation of individual risk factors. This study represents the first ML-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1231 - 1236
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0169

based approach to assess comprehensive risk utilizing EMR data from a large regional healthcare system. The implications of this study's findings include the potential to minimize the necessity for additional diagnostic tests and provide a cost-effective screening strategy.

Conclusion

In conclusion, our study demonstrates the superiority of Machine Learning (ML) models over the traditional PCE Risk Calculator in predicting Atherosclerotic Cardiovascular Disease (ASCVD) risk. By leveraging advanced ML techniques and Electronic Medical Record (EMR) data, we achieved higher prediction accuracy and efficiency. Incorporating both Longitudinal (LT) and Cross-Sectional (CS) features significantly enhanced predictive performance, reducing the need for patient screening while maintaining high sensitivity levels. Our Random Forest (RF) model, particularly when utilizing combined CS and LT characteristics (RF-LTC), achieved an impressive Area Under Curve (AUC) score of 0.902. Notably, our model identified 90% of positive ASCVD cases with screening of only 43% of patients, outperforming the PCE Risk Calculator which required screening 69% of patients for similar accuracy. These findings underscore the potential of ML-driven approaches in optimizing preventive care strategies and resource allocation in combating cardiovascular diseases. Further validation and implementation of our model hold promise for improving ASCVD risk assessment and informing targeted interventions, ultimately contributing to better patient outcomes and healthcare efficiency.

Acknowledgements

I express my sincere thanks to my guide Dr. Nilesh Ashok Suryawanshi. The faith & confidence shown by her in me, boosted me and motivated me to perform better.

References

[1]. S.S. Virani, A. Alonso, H.J. Aparicio, E.J. Benjamin, M.S. Bittencourt, C. W. Callaway, A.P. Carson, A.M. Chamberlain, S. Cheng, F.N. Delling, M.S.V. Elkind, K.R. Evenson, J.F. Ferguson, D.K. Gupta, S.S. Khan, B.M. Kissela, K.L. Knutson, C., D. Lee, T.T. Lewis, J. Liu, M.S. Loop, P.L. Lutsey, J. Ma, J. Mackey, S.S. Martin, D., Matchar, M.E. Mussolino, S.D. Navaneethan, A.M. Perak, G.A. Roth, Z. Samad, G.M. Satou, E.B. Schroeder, S.H. Shah, C.M. Shay, A. Stokes, L.B. VanWagner, N.-Y. Wang, C.W. Tsao, Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association, Circulation 143 (8) (2021), https://doi.org/10.1161/CIR.0000000000000950.

[2]. D.M. Lloyd-Jones, L.T. Braun, C.E. Ndumele, S.C. Smith, L.S. Sperling, S.S. Virani, R.S. Blumenthal, Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report from the American Heart Association and American College of Cardiology, Circulation 139 (25) (2019), https://doi.org/10.1161/CIR.0000000000000638.

[3]. Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD., Risk scoring for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2017;3:CD006887. Epub 2017/03/16. doi: 10.1002/14651858.CD006887.pub4. PubMed PMID: 28290160; PubMed Central PMCID: PMCPMC6464686.

[4]. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. Circulation. 2018: CIR0000000000000624. Epub 2018/12/20. doi: 10.1161/CIR.0000000000000624. PubMed PMID: 30565953.

[5]. Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography

angiography: analysis from the CONFIRM registry. Eur Heart J. 2020;41(3):359-67. Epub 2019/09/13. doi: 10.1093/eurheartj/ehz565. PubMed PMID: 31513271; PubMed Central PMCID: PMCPMC7849944.

[6]. K.M. Chinnaiyan, P. Peyser, T. Goraya, K. Ananthasubramaniam, M. Gallagher, A. DePetris, J.A. Boura, E. Kazerooni, C. Poopat, M. Al-Mallah, S. Saba, S. Patel, S. Girard, T. Song, D. Share, G. Raff, Impact of a Continuous Quality Improvement Initiative on Appropriate Use of Coronary Computed Tomography Angiography, J. Am. Coll. Cardiol. 60 (13) (2012) 1185–1191.

[7]. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Sr., Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63(25 Pt B):2935-59. Epub 2013/11/ 19. doi: 10.1016/j.jacc.2013.11.005. PubMed PMID: 24239921; PubMed Central PMCID: PMCPMC4700825.

[8]. R. Nakanishi, P.J. Slomka, R. Rios, J. Betancur, M.J. Blaha, K. Nasir, M. D. Miedema, J.A. Rumberger, H. Gransar, L.J. Shaw, A. Rozanski, M.J. Budoff, D. S. Berman, Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths, JACC Cardiovasc Imaging. 14 (3) (2021) 615–625, https://doi.org/10.1016/j.jcmg.2020.08.024

[9]. M. Kavousi, M.J.G. Leening, D. Nanchen, P. Greenland, I.M. Graham, E. W. Steyerberg, M.A. Ikram, B.H. Stricker, A. Hofman, O.H. Franco, Comparison of Application of the ACC/AHA Guidelines, Adult Treatment Panel III Guidelines, and European Society of Cardiology Guidelines for Cardiovascular Disease Prevention in a European Cohort, JAMA 311 (14) (2014) 1416, https://doi.org/10.1001/ jama.2014.2632.

[10]. Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. J Am Coll Cardiol. 2016;67(18):2118-30. Epub 2016/05/07. doi: 10.1016/j.jacc.2016.02.055. PubMed PMID: 27151343; PubMed Central PMCID:

[11]. Ward, A. Sarraju, S. Chung, J. Li, R. Harrington, P. Heidenreich, L. Palaniappan, D. Scheinker, F. Rodriguez, Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population, npj Digit. Med. 3 (1) (2020), ttps://doi.org/10.1038/s41746-020-00331-1.

[12]. A. Ward, A. Sarraju, S. Chung, J. Li, R. Harrington, P. Heidenreich, L. Palaniappan, D. Scheinker, F. Rodriguez, Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population, npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00331-1.

[13]. Lundberg SM, Lee S-I, editors. A Unified Approach to Interpreting Model Predictions. NIPS; 2017.