# Health Information Technology in Cardiovascular Disease Prediction: Advancements and Challenges

Priya Kaur, Pavan Marturu and Akinola Suliat

## I. INTRODUCTION

Cardiovascular disease (CVD) is a primary source of morbidity and mortality worldwide with 17.9 million deaths per year, it is no longer thought of as a disease of the elderly. According to recent studies, young adults are also at risk of having CVD. The prevention or delay of CVD onset can be achieved through early detection and management of CVD risk factors, which also improves health outcomes. In recent years, CVD risk factor management and prediction have benefited greatly from the development of health informatics.

Various risk factors, such as age, gender, body mass index (BMI), blood pressure, cholesterol levels, smoking habits, family history, and levels of physical activity, are examined to predict CVD in young adults. The gathering, storage, processing, and analysis of substantial amounts of data pertaining to these risk factors are made possible by the usage of health informatics architecture. Wearables, mobile applications, electronic health records, and other sources can all provide this information.

Population health information architectures have emerged as critical tool for improving health outcomes and reducing healthcare costs. These architectures are aid in the collection, storage, analysis, and effective use of health-related data. This can be used to identify high-risk populations and make informed clinical decision-making. However, the use of population health information architectures for predicting and preventing chronic diseases such as cardiovascular disease has been limited by the complexity and heterogeneity of health data.

The use of machine learning techniques in the health information architecture has helped mitigate some of the challenges. Machine learning algorithms can analyze large and complex datasets to identify patterns and relationships that are usually missed by traditional statistical methods. Hence it can be said that these techniques hold great potential for improving the accuracy of risk prediction models for chronic diseases like coronary heart disease.

The health informatics architecture will have multiple stages to predict the risk of CVD in young adults. First, information will be gathered via electronic health records, which will also include demographic data, medical history, lifestyle factors, and test results from labs. To clean and standardize the data, pre-processing will be done. The most pertinent variables for predicting the risk of CVD will subsequently be found using feature selection approaches.

Machine learning algorithms including logistic regression, random forests, and neural networks will be utilized to train predictive models utilizing the chosen features. To confirm the models' accuracy and generalizability, cross-validation methods will be used for validation.

In a clinical environment, the prediction models will be used as a decision support system. The technology will be included in the electronic health record system and will give healthcare practitioners risk scores and suggestions. Personalized risk assessments and treatment plans will also be generated by the system for each patient based on their unique risk variables.

The study population consisted of 251,791 patients with hypertension from the Shenzhen Health Information platform, Electronic Health Record database. The patients were aged 18-90 years and had at least two outpatient visits or one inpatient visit for hypertension between January 1, 2010, and December 31, 2018.

Overall, in this paper highlights the importance of using robust databases, standardized coding systems, and appropriate data processing techniques in leveraging electronic health records and machine learning algorithms for improving patient outcomes.

## II. LITERATURE REVIEW

The purpose of this literature review is to provide an overview of the work that has been done on this topic. Du et al. (2020) used electronic health records and big data to develop and evaluate a machine-learning model for the prediction of coronary heart disease in patients with hypertension. The study uses several machine learning models such as decision trees, random forests, support vector machines, and artificial neural networks. They have even used feature selection to identify the most important variables for the prediction of cardiovascular disease. The model has high accuracy and demonstrates the potential of machine learning in predicting CVD/CVS risk in a complex population, I.e., a population with co-morbidities.

Alaa et al. (2019) conducted a study on over 400,000 UK Biobank participants to demonstrate the feasibility and potential of automated machine learning for cardiovascular disease risk prediction. Several machine algorithms such as logistic regression, neural networks, and gradient boosting have been used in this study. This study shows the potential for using machine learning models on a large-scale population health architecture. This study puts emphasis on feature selection, model calibration, and interoperability for the development of highly accurate models.

Dinh et al. (2019) developed a machine learning model using logistic regression, decision tree, random forest, and gradient boosting for predicting diabetes and cardiovascular disease, showing promising results in terms of accuracy and predictive performance.

Goldstein et al. (2017) conducted an in-depth review of how machine learning techniques can overcome the drawbacks that one faces with using traditional regression techniques for cardiovascular risk prediction. The study focused on selecting appropriate features, optimizing the model performance with various techniques such as hyperparameter tuning, and mitigating issues such as class imbalance and data quality would give us more ore accurate results.

Quesada et al. (2019) developed a machine learning model for predicting cardiovascular risk using demographic, clinical, and laboratory data from electronic health records. This study focuses on the potential of convolutional neural networks (CNN) and recurrent neural networks in predicting of CVD. The study demonstrates the potential of machine learning models in improving the accuracy of predicting the risk of cardiovascular disease when compared to traditional risk assessment tools such as the Framingham Risk Score.

There are some instances where machine learning models have been shown to be effective to predict the outcomes of cardiovascular diseases such as heart failure. Shah et al. (2017) developed a machine learning model for predicting heart failure in patients with diabetes. The study's results are that the model can differentiate and identify patients at substantial risk for heart failure.

In addition to improving risk prediction, machine learning techniques have also been used to identify new risk factors and potential targets for intervention. For example, Kohane et al. (2017) used machine learning to identify novel risk factors for heart failure, including low serum calcium levels and high diastolic blood pressure.

### A. *limitations*

While machine learning techniques have shown great promise in predicting cardiovascular disease, there are also limitations and challenges that need to be addressed.

One of the limitations is the need for high-quality data. Du et al. (2020) used electronic health records to develop their machine-learning model for predicting coronary heart disease in patients with hypertension. However, the quality of electronic health records is not consistent. There could be missing data or inaccuracies that can limit the effectiveness of machine learning models.

Another challenge is the need for interpretable models. While machine learning models can be highly accurate, they can also be difficult to interpret which can limit their usefulness for clinical decision-making because clinicians might be hesitant to use models that they do not fully comprehend. In the study by Alaa et al. (2019), the importance of interpretability in developing machine learning models for cardiovascular risk prediction has been emphasized.

Another limitation is the potential for bias in machine learning models. Biases can arise from many sources, including the data used to train the models and the features selected for inclusion. These biases lead to models that are not entirely representative of the underlying populations. In the study by Quesada et al. (2019), it was noted that the potential sources of bias need to be considered when developing machine learning models for cardiovascular risk prediction.

One of the challenges that is most talked about in Health information is security and privacy. Health data is subject to strict laws and regulations such as HIPPA. Dinh et al. (2019) in his study noted that it is important to comply with the regulations and guidelines regarding data security and privacy while building machine learning models for predicting diseases.

Finally, there are also challenges related to the architecture and infrastructure needed to support machine learning applications in healthcare. With the ever-growing volume and complexity of healthcare data, there is need for scalable healthcare information architecture to support machine learning applications. In the study by Kohane et al. (2017), it was emphasized that it is important to develop flexible and scalable infrastructure to support machine learning applications in healthcare.

### B. *Potential Gaps*

The potential gaps in research are that most studies are data sources diversity. Most studies are using data from either a single population or a particular healthcare system. This leads to a limited diversity in the study population hence limiting the ability of the model to generalize new data or varied data. Another gap is the lack of external validation performed on an independent dataset. Some of the studies only used internal validation which might lead to overfitting and again limit the ability of the model to adapt and perform well with different populations and settings. There is limited discussion on the clinical implications of the model and how it can be integrated into the current clinical workflow to improve patient care outcomes. Exploring further on the above point, these studies fall short on the limited considerations of clinical factors while building these models. Most studies take only the traditional clinical risk factors into consideration. There is no mention of how changes in clinical management or patient behavior can impact risk prediction. Most of these studies do not discuss the ethical and legal implications of using machine learning to predict heart diseases.

Overall, these studies demonstrate the potential of machine learning techniques for improving cardiovascular disease risk prediction, identifying new risk factors, and informing clinical decision-making. However, there are also challenges and limitations to be addressed, including data quality and availability, model interpretability, and the need for rigorous validation and evaluation of machine learning models in real-world settings.

### C. *Conflicting Evidence*

A study conducted by Bae et al. (2020) concluded that machine learning models are less accurate when compared to traditional risk prediction models such as the Framingham

Risk score in predicting cardiovascular disease. Another study by Wang et al. (2020) showed that machine learning models usually, due to their training nature, overestimate the results in low-risk populations which leads to unnecessary interventions. A study by Attia et al. (2019) touches on the bias nature of machine learning models. The study found that models that were trained on a predominantly white population runs a risk of overestimating the risk of CVD in as more diverse population which would lead to wrong diagnosis. This topic is further studied in a study by Khera et al (2019) which concluded that ML models have the potential to predict CVD accurately when trained on a larger dataset but tend to give false results when trained on smaller datasets. A study by Lipton et al. (2018) concluded that the black -box nature of machine learning makes it hard to trust the model completely and raises questions about its biased nature. Machine learning models should be used in conjunction with traditional tools was concluded by a study conducted by Pencina et al. (2019).

## III. CASE STUDY

The aim of the study was to predict the onset of coronary heart disease (CHD) in patients with hypertension using electronic health record (EHR) data. The researchers collected EHR data from the Shenzhen Health Information platform, which contained the clinical records of 83 local public hospitals and over 600 community health service centers from 2010 to 2018. The data were de-identified by the platform administrators under the supervision of the Shenzhen Municipal Health Commission before being collected for the study. A total of 251,791 registered patients with hypertension were identified in the platform data. The collected EHR data for each patient included regular chronic disease follow-up records, inpatient and outpatient records, and clinical examinations and biochemical tests. The CHD diagnosis results were extracted from the main diagnosis field of the inpatient or outpatient records using the International Statistical Classification of Diseases and Related Health Problems (ICD)-10 diagnostic codes I20 to I25 or the keywords related to CHD conditions.

### A. Feature Processing

The study in question involved a unique dataset that contained multiple records with different record times for each patient. Thus, data pre-processing and feature variable extraction, selection, and construction were critical steps for establishing and analysing the model. To begin with, variables with over 20 percent missing values were excluded from the study. Next, the researchers performed text parsing to convert diagnostic results, which were a combination of ICD codes and natural language text input, into corresponding ICD codes. If ICD codes were available, they were used directly as features of the samples. However, if not, a rule-based, in-house-designed lexical parsing code was employed, which used keyword mapping and error corrections to map each ICD code item to different texts through a regular expression of keywords. The parsing procedure was iterative, and at the end of each loop, unparsed texts were collected, sorted by word frequency, and manually inspected. The expressions were then modified to match more text, including tolerating typographical errors, until the unparsed texts were considered non-informative.

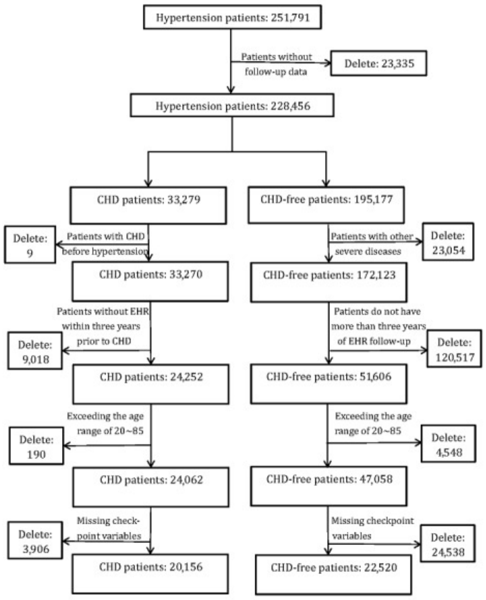

Hypertension patients: 251,791
Patients without follow-up data
Delete: 23,335
Hypertension patients: 228,456
CHD patients: 33,279
CHD-free patients: 195,177
Delete: 9
Patients with CHD before hypertension
Patients with other severe diseases
Delete: 23,054
CHD patients: 33,270
CHD-free patients: 172,123
Patients without EHR within three years prior to CHD
Delete: 9,018
Patients do not have more than three years of EHR follow-up
Delete: 120,517
CHD patients: 24,252
CHD-free patients: 51,606
Exceeding the age range of 20~85
Delete: 190
Exceeding the age range of 20~85
Delete: 4,548
CHD patients: 24,062
CHD-free patients: 47,058
Missing check-point variables
Delete: 3,906
Missing checkpoint variables
Delete: 24,538
CHD patients: 20,156
CHD-free patients: 22,520

Fig. 1. Patient cohort data processing. CHD: coronary heart disease; EHR: electronic health record.

Accounting was then carried out by gathering features from multiple sources representing the same physiology index, such as examination, inpatient, and outpatient records, and calculating their maximum, minimum, or average values to create new features. To avoid sparsity in value distribution, rare diagnostic symptoms and similar symptoms were merged into a single variable. Finally, the researchers divided the follow-up period of each patient into early and late halves at the mid-time points. They then accounted for the frequency of specified events such as in-hospital or out-hospital visits and symptom onset for each half and used the ratios as a new variable representing the trending status of the patients.

### B. Machine-Learning Algorithms

Machine learning is a widely used technique in many fields, including healthcare. There are several algorithms used in machine learning, and one of them is XGBoost. XGBoost is an integrated machine learning algorithm that uses multiple decision trees with gradient boosting as the framework. The algorithm minimizes a differentiable convex loss function that measures the difference between the prediction and the target. In addition, it uses a regularization term to penalize the complexity of the model to avoid overfitting. XGBoost also supports missing values and sparse feature format, making it easy to feed data as a sparse matrix.

Another algorithm commonly used in machine learning is Support Vector Machine (SVM). SVM is a supervised learning algorithm that uses a maximum-margin hyperplane to classify data. It trains a function that calculates a score for a new input to separate samples into two classes by building this hyperplane. Logistic regression is also a popular machine learning algorithm used for classification. It is a generalized linear regression analysis model that assumes the data follows the Bernoulli distribution. The algorithm uses the method of maximizing the likelihood function to solve the parameters with gradient descent to achieve the purpose of classifying the data.

The decision tree algorithm is another commonly used machine learning algorithm. It builds a model based on the characteristics of data using a tree structure. The process of constructing decision trees usually includes feature selection, tree generation, and pruning. The K-Nearest Neighbor (KNN) algorithm is used to classify data by comparing the characteristics of the input test data with the corresponding features of the training set. The algorithm finds the top K dataset most similar in the training set and summarizes the most frequently occurring classification among the K most similar data to classify the test data.

Random forest is an integrated learning algorithm that integrates multiple decision trees into a single classifier. The algorithm selects different splitting features and training samples to generate a forest of a large number of decision trees. When predicting unknown samples, each tree in the forest makes a decision, which improves the accuracy of the prediction compared to a single decision tree. By statistically determining the results of the decision, the classification with the highest number of votes is taken as the final classification result.

### C. Missing Data

XGBoost handles missing values differently than other algorithms. Instead of filling in missing values with an average value, XGBoost marks missing values as "missing" and only uses non-missing samples to create trees. This means that XGBoost does not need to impute missing values and can still build accurate models.

### D. Evaluation Criteria

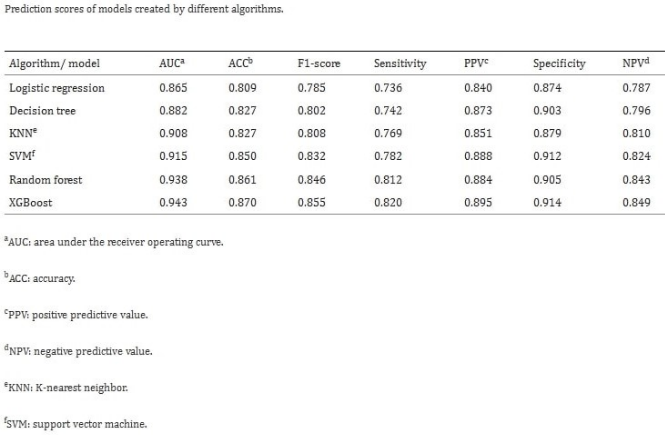
The performance of a classification model can be evaluated using a confusion matrix, which contains information about the number of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP) predicted by the model. Various performance indices can be computed from the confusion matrix, such as accuracy, sensitivity, positive predictive value (PPV), specificity, negative predictive value (NPV), and F1-score. Accuracy is calculated as the proportion of correctly classified samples to the total number of samples. Sensitivity measures the percentage of TP samples that were correctly predicted, while PPV measures the percentage of positive samples that were predicted correctly. Specificity measures the proportion of TN samples that were correctly predicted, while NPV measures the percentage of negative samples that were predicted correctly. The F1-score is calculated as a harmonic average of model accuracy and recall, and is used as an overall measure of the model's performance.

To further evaluate the model, a receiver operating characteristic (ROC) curve can be plotted by sorting the samples based on the prediction results of the model and predicting the samples as positive examples one by one. The ROC curve displays the trade-off between the true positive rate (TPR) and false positive rate (FPR) for different classification thresholds. The area under the ROC curve (AUC) is then selected as the main evaluation index, with a larger AUC value indicating a better classification result and prediction effect of the model. The AUC is calculated as the sum of the ranks of the positive examples divided by the product of the total number of positive and negative examples.

Overall, these performance indices and evaluation methods provide a comprehensive assessment of the model's ability to classify samples accurately and effectively, which is important for a wide range of applications in machine learning and data analysis.

### E. Results

*1) Model Prediction Performances:* The researchers used a dataset with 65 feature variables to train and test several machine-learning models for predicting the risk of coronary heart disease (CHD). They found that the XGBoost ensemble method had the highest accuracy on the test dataset, followed closely by the random forest method. These nonlinear models outperformed the linear logistic regression model commonly used in previous risk prediction models.



Prediction scores of models created by different algorithms.

| Algorithm/ model | AUC[a] | ACC[b] | F1-score | Sensitivity | PPV[c] | Specificity | NPV[d] |
|---|---|---|---|---|---|---|---|
| Logistic regression | 0.865 | 0.809 | 0.785 | 0.736 | 0.840 | 0.874 | 0.787 |
| Decision tree | 0.882 | 0.827 | 0.802 | 0.742 | 0.873 | 0.903 | 0.796 |
| KNN[e] | 0.908 | 0.827 | 0.808 | 0.769 | 0.851 | 0.879 | 0.810 |
| SVM[f] | 0.915 | 0.850 | 0.832 | 0.782 | 0.888 | 0.912 | 0.824 |
| Random forest | 0.938 | 0.861 | 0.846 | 0.812 | 0.884 | 0.905 | 0.843 |
| XGBoost | 0.943 | 0.870 | 0.855 | 0.820 | 0.895 | 0.914 | 0.849 |

[a] AUC: area under the receiver operating curve.

[b] ACC: accuracy.

[c] PPV: positive predictive value.

[d] NPV: negative predictive value.

[e] KNN: K-nearest neighbor.

[f] SVM: support vector machine.

Fig. 2. Prediction scores of models created by different algorithms.

To address concerns about the possible bias introduced by variations in the total follow-up time of patients, the researchers divided the test sets into two groups based on whether CHD onset occurred within 3 years and the total follow-up time. They applied the same prediction model on both groups separately and found that the performance of the model was similar with no statistically significant difference between the risk score distributions. This suggests that the inclusion of CHD patients with less than 3 years of follow-up time did not introduce observable data bias and the models developed were reliable for generalization.

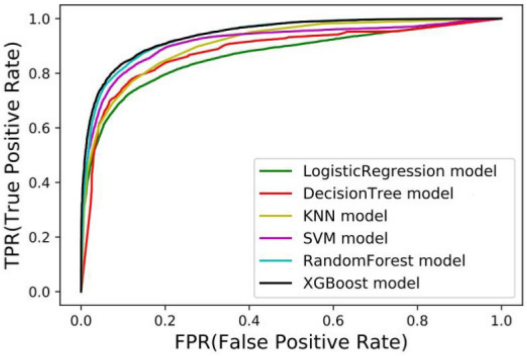


Fig. 3. The receiver operating characteristic curves of models established by different algorithms. AUC: area under the curve

*2) Impact of Population Size on Model Performance:* An experiment was conducted to investigate the impact of population size on the accuracy and reliability of disease prediction models. The experiment utilized various subpopulations with fixed variable sets of different sizes. The results,indicate that an increase in population size leads to improved model prediction performance, given the characteristic variables are fixed. However, with an adequate variable set, even a small training population size can produce fairly competitive performance (AUC>0.8), surpassing the results obtained with a large training population but limited feature variables. Therefore, the findings suggest that while population size is an important factor in building disease risk prediction models, it is not an overwhelming limitation.

## IV. RESULTS AND DISCUSSION

Implementing machine learning models into the current healthcare system comes with a substantial cost. The cost varies based on several factors such as the complexity of the model, the size of the dataset, resources required for training, deployment, and continuous evaluation or upkeep.

The cost of implementing different machine learning models can vary widely depending on several factors, including the complexity of the model, the size of the dataset, and the resources required for training and deployment. Such sophisticated models require more computational resources which eventually leads to a higher cost of training. This point is touched on by a study conducted by Zadrozny et al. (2020) which concluded that training a deep learning model takes several weeks to train on a single CPU. Thus, the cost of additional specialized hardware such as GPUs comes into the picture. GPUs can cost upwards of thousands of dollars. Additionally, servers with high end CPUs are required to handle the computational requirements of operating large scale machine learning models. The cost factor is influenced by training time as well. The time required to train a model varies depending on the complexity of the model as well as the size of the dataset. This indirectly influences the total cost factor. Now coming to one of the major factors-deployment of the models. Deployment of the model comes with substantial costs such as the cost of hosting, cost of maintenance, infrastructure cost, and cost of training the staff. Moreover, these models need periodic updates to make sure that the model is up-to-date and accurate, which adds up to the cost. Some machine learning models tend to be more costly than others. For example, models such as decision tree may be relatively less expensive when compared to convulational neural networks.

A paper by Goldstein et al. (2017) studied and compared the cost and performance of various machine learning models for CVD prediction using data from electronic health records.

The paper gave an example of cost of implementing machine learning models for CVD risk prediction depending on specific algorithm and dataset used. The cost of training a random forest model was around 1.46 dollars per patient, and on the other hand cost of training a deep learning model was around 5.70 dollars per patient. The paper also states that this cost might go up when considering the large number of patients that needs to be analyzed in most healthcare settings. Another paper that studies the cost is by Goldstein et al. (2019). This paper reports that the total cost of running the model was around 250,000 USD. Deepheart project that is conducted by researchers from University of California, San Francisco, and Cardiogram to use deep learning algorithm to analyze data from earable devices to predict risk of CVD is reported to cost around 2 million USD.

Coming to model efficiency, the model developed by Du et al. (2020) using electronic health records and big data achieved an Area Under the Curve (AUC) of 0.906. Model developed by Quesada et al. (2019) using a dataset of 162,424 patients achieved an AUC of 0.801. Goldstein et al (2017) compared different machine learning models on a dataset of 999,983 patients and found that random forest is the best model that achieved an AUC of 0.836. Alaa et al. (2019) developed a model using data of 423,604 from UK Biobank got an AUC of 0.853. Using dataset of 38,026 patients, model developed by Dinh et al. (2019) achieved an AUC of 0.76.

Overall, the models used in these studies outperformed the traditional regression based models. These models are important tools for population health management since they help identify high-risk CVD individuals who could benefit from early interventions.

Population health management effectiveness is analyzed by the performance of the models at identifying the high risk individuals. It can also be measured by analyzing the predicted and observed incidence of CVD in the population tested. The cost effectiveness can be studied by considering the economic impact that accurate detection and hence early intervention played in the population.

However, it is important to note that the machine learning models are only one component of an effective population health management strategy. Along with these models to accurately identify the high risk individuals, effective interventions need to be put into place to prevent CVD. Additionally, it is important to consider the ethical implications of using machine learning models for risk prediction.

Overall, the studies demonstrate the potential of machine learning models for predicting CVD risk using EHRs and other health data sources. The effectiveness of these population health management tools depend on a range of factors, including but not limited to the quality of the underlying data and interventions.

## V. Conclusion

Considering the findings from many studies mentioned in this article, it can be concluded that the use of machine learning models for predicting cardiovascular disease risk has promising results. The studies show that machine learning has the potential to accurately predict cardiovascular risk using electronic health records and big data.

The studies also concluded that machine learning models are better at predicting CVD risk when compared to traditional regression models. They also address the challenges one might face during analysis of risk prediction.

High quality and standardized data can further improve the architecture for machine learning models in the prediction of cardiovascular disease. Transparency and interpretability of the models should also be taken into consideration while building these models. Continuous monitoring and evaluation is necessary to maintain the effectiveness and accuracy of the models.

This can be achieved by including a more diverse range of population to train and test the model. This would ensure that the model can adjust and perform well with generalized data.

Another way could be to include data in addition to electronic health records. Data such as genetic information, lifestyle factors, and environmental factors can add more layers to the model thus improving the results of the prediction.

Data cleaning and data pre-processing should be an essential part of building models to ensure that the data is of good quality. Lastly, the model should be integrated well into a clinical setting.

While these recommendations look feasible, there are certain steps that need to be taken to effectively implement them. Processing and cleaning data to ensure data is standardized and of great quality takes a lot of time and monetary resources. Ensuring that the models are built in a way that they can be easily understood and used by healthcare workers while not compromising on quality takes effort.

Additionally, it is important to ensure that these models can be used in a safe, responsible, and ethical way. Further research is needed to explore the real life application of machine learning applications in clinical settings to predict and work in the conjugation of a healthcare worker.

Overall, the possible real-world applicability of such models can be in the form of a web-based tool that assesses disease risk through a survey questionnaire. The participants can choose to consult a doctor based on the score received. Before this becomes a reality in a healthcare setting, we need to make sure that the computational capacity in healthcare systems improves.

## References

[1] Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, Fan J, Cheng W, Chen XH, Cai Y. Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. JMIR Med Inform. 2020 Jul 6;8(7):e17257. doi: 10.2196/17257. PMID: 32628616; PMCID: PMC7381262.

[2] Quesada JA, Lopez-Pineda A, Gil-Guillén VF, Durazo-Arvizu R, Orozco-Beltrán D, López-Domenech A, Carratalá-Munuera C. Machine learning to predict cardiovascular risk. Int J Clin Pract. 2019 Oct;73(10):e13389. doi: 10.1111/ijcp.13389. Epub 2019 Aug 4. PMID: 31264310.

[3] Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017 Jun 14;38(23):1805-1814. doi: 10.1093/eurheartj/ehw302. PMID: 27436868; PMCID: PMC5837244.

[4] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174944 Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019 May 15;14(5):e0213653. doi: 10.1371/journal.pone.0213653. PMID: 31091238; PMCID: PMC6519796.

[5] Dinh, A., Miertschin, S., Young, A. et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19, 211 (2019). https://doi.org/10.1186/s12911-019-0918-5

[6] Nakanishi R, Slomka P, Rios R, et al. Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths. J Am Coll Cardiol Img. 2021 Mar, 14 (3) 615–625.https://doi.org/10.1016/j.jcmg.2020.08.024

[7] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12(4): e0174944. https://doi.org/10.1371/journal.pone.0174944

[8] Stephan Dreiseitl, Lucila Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics, Volume 35, Issues 5–6, 2002, Pages 352-359, ISSN 1532-0464, https://doi.org/10.1016/S1532-0464(03)00034-0.

[9] Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang, A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data, Medical Engineering  Physics, Volume 105, 2022, 103825, ISSN 1350-4533, https://doi.org/10.1016/j.medengphy.2022.103825.

[10] 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines.

[11] Eichner J, et al. (2019). Data quality in electronic healthcare records: An overview of quality domains and respective methods. Journal of Medical Systems, 43(9), 286.

[12] Wang Y, et al. (2019). Big data analytics for hypertension and coronary heart disease: Development and validation of a model using machine learning methods with electronic health records. Journal of Medical Internet Research.