HHS Public Access

Author manuscript

Am Heart J Plus. Author manuscript; available in PMC 2023 February 13.

Published in final edited form as:

Am Heart J Plus. 2022 December ; 24: . doi:10.1016/j.ahjo.2022.100232.

Short-term prediction of coronary artery disease using serum metabolomic patterns

Ben Omega Petrazzini[a,b], **Akhil Vaid**[a,c,d], **Joshua K. Park**[a,b,e], **Carla Marquez-Luna**[a,b], **Ha My Vy**[a,b], **Aparna Saha**[a,c,d], **Kumardeep Chaudhary**[a,b], **Judy Cho**[a,b], **Lili Chan**[a,c,d], **Edgar Argulian**[a,c], **Jagat Narula**[a,c], **Girish Nadkarni**[a,c,d,f,*,1], **Ron Do**[a,b,d,*,1]

[a]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[b]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[c]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[d]The BioMe Phenomics Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[e]Medical Scientist Training Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[f]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Keywords

Coronary artery disease prediction; Serum metabolic patterns; Machine learning; Pooled cohort equations; Cardiovascular disease prevention; Personalised medicine

Current risk assessment methods for coronary artery disease (CAD), such as the pooled cohort equations (PCE) [1], are based on a limited combination of known cardiovascular risk factors. Studies have shown that alternative forms of data such as electrocardiogram [2] and clinical data in electronic health records [3] can improve CAD prediction beyond that of the PCE. These studies suggest that, compared to traditional risk assessment models based on traditional risk factors, alternative sources of data can better capture the entire representation of disease risk, resulting in greater prediction power.

Metabolomics data yields relevant biological information of an individual's health status, independent of clinical records. This additional source of information can capture novel evidence of CAD susceptibility, not considered by current approaches based on a reduced

*Corresponding authors at: Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, PO Box 1243, New York, NY 10029, USA. girish.nadkarni@mountsinai.org (G. Nadkarni), ron.do@mssm.edu (R. Do).

[1]These authors jointly supervised this work.

CRediT authorship contribution statement

Drs. Nadkarni and Do jointly supervise this work.

number of traditional risk factors. Hence, a prediction score based on metabolomic features carries significant potential to advance CAD risk assessment. In this study, we evaluate whether serum metabolomic patterns integrated in a machine learning (ML) framework can accurately predict CAD one year ahead of diagnosis in a multiethnic cohort comprised of 251 CAD cases and 1603 controls. We compared this model to the conventional clinical risk score, the PCE.

Blood samples from 2000 unrelated individuals from the Bio*Me* Biobank were processed to obtain metabolomic data using the Metabolon assay. A total of 1654 available features were considered. Age, gender and self-reported ethnicity were included as covariates. We considered data one year prior to CAD diagnosis. Individuals with more than 60 % missing values were discarded resulting in 1854 total individuals available for posterior analyses. Missing data for the remainder were imputed using a random forest-based approach.

A ML approach was used to develop the metabolomics model for CAD prediction (Fig. 1A). To minimize overfitting, the workflow was replicated 100 times using different samples for training and testing in each iteration. The training set was built by randomly selecting 80 % of cases and an equal number of controls. All subsequent steps were performed only on the train set and then applied to the test set. Feature selection was performed in each iteration using a wrapper approach built around a random forest algorithm. This reduces the complexity of the model, making the prediction clinically interpretable. Non-selected features were then removed from the test set accordingly. Continuous features were scaled within the train set and the resulting metrics were used to scale the test set accordingly. The model regresses predictions from three algorithms to compute a final score, namely random forest, gradient boosted trees and support vector machine with polynomial kernel. For each algorithm, hyperparameters were optimised using an internal 10-fold cross-validation within the train set. The resulting model was then used to predict CAD in the test set. Performance metrices representative of the entire population were obtained by randomly selecting samples in each iteration while still avoiding overfitting. The reported metrices correspond to the mean and standard deviation across 100 iterations.

Standard performance metrics, namely sensitivity (recall), specificity, accuracy, area under receiver operating characteristic (AUROC) curve and standard error (SE) were used to assess the performance of each model.

A total of 1654 metabolomic features were obtained for 986 African Americans, 740 Hispanic Americans, 83 European Americans and 45 individuals from other ancestries. Compared to the PCE, the metabolomic model improved CAD discrimination power by 15 % (AUROC = 0.66, SE = 0.04 for PCE vs. AUROC = 0.81, SE = 0.03 for metabolomic) in a blind test set (Fig. 1B). It identifies 75 % of CAD cases one-year prior to diagnosis (Sensitivity = 0.75, SE = 0.05) with high confidence (PPV = 0.73, SE = 0.03) (Fig. 1C). A predictive model using both metabolomic information and the PCE (metabolomic + PCE) carries slightly stronger discriminative power (AUROC = 0.82, SE = 0.03) (Fig. 1B). Furthermore, a feature importance analysis identifies known metabolites relevant to cardiovascular pathophysiology including creatinine, alanine, aspartate and benzoate

metabolism [4] (Fig. 1D). Metabolomic features not yet mapped to a biological process are labelled “Unnamed”.

These results suggest that metabolomic information can be used to determine one-year risk estimations for CAD. Importantly, both the discriminative and case prediction power of a metabolomic model is superior to the PCE [5], indicating it has clinical applicability and utility. Furthermore, by using out-of-hospital information, it can complement PCE-based predictions to identify high-risk individuals that are not captured by clinical data in electronic health records.

This study has some limitations. First, the study does not include external validation due to a lack of available metabolomics data in an external cohort. Therefore, we were not able to assess the model performance in a validation cohort. However, to minimize overfitting originating from sampling biases, 100 iterations of training and testing on independent datasets were performed. Second, a health system-based cohort (Bio*Me*) was used even though the metabolomic model has wider utility beyond health system settings. Of note, this was a multiethnic cohort with the advantage of having clinically validated CAD diagnoses. Finally, the PCE used here to assess short-term prediction of CAD had been developed originally for long-term 10-year risk assessment. We show in a previous study that the PCE can predict short-term risk for CAD [3]. In conclusion, this study shows that metabolomic data carries sufficient predictive power to discriminate CAD cases using a simple metabolomics blood assay.

Source of funding

Dr. Do is supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) (R35-GM124836) and the National Heart, Lung, and Blood Institute of the NIH (R01-HL139865 and R01-HL155915). Dr. Nadkarni is supported by NIH grants R01-DK108803 and R01-HL155915.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr. Nadkarni reported being a scientific co-founder, consultant, advisory board member, and equity owner of Renalytix AI, a scientific co-founder and equity holder for Pensieve Health, a consultant for Variant Bio, and receiving grants from Goldfinch Bio and personal fees from Renalytix AI, BioVie, Reata, AstraZeneca, and GLG Consulting. Dr. Do reported receiving grants from AstraZeneca, grants and non-financial support from Goldfinch Bio, being a scientific co-founder, consultant and equity holder for Pensieve Health, and being a consultant for Variant Bio. The remaining authors have nothing to disclose.

References

[1]. Goff DC, Lloyd-Jones DM, Bennett G, et al. , 2013 ACC/AHA guideline on the assessment of cardiovascular risk, Circulation 129 (25) (2013) S49–S73. [PubMed: 24222018]

[2]. Acharya UR, Fujita H, Lih OS, Adam M, Tan JH, Chua CK, Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network, Knowl.-Based Syst 132 (2017) 62–71.

[3]. Petrazzini BO, Chaudhary K, Márquez-Luna C, et al. , Coronary risk estimation based on clinical data in electronic health records, J. Am. Coll. Cardiol 79 (12) (2022) 1155–1166. [PubMed: 35331410]

[4]. Fromentin S, Forslund SK, Chechi K, et al. , Microbiome and metabolome features of the cardiometabolic disease spectrum, Nat. Med 28 (2) (2022) 303–314. [PubMed: 35177860]

[5]. Rana JS, Tabada GH, Solomon MD, Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population, J. Am. Coll. Cardiol 67 (18) (2016) 2118–2130. [PubMed: 27151343]

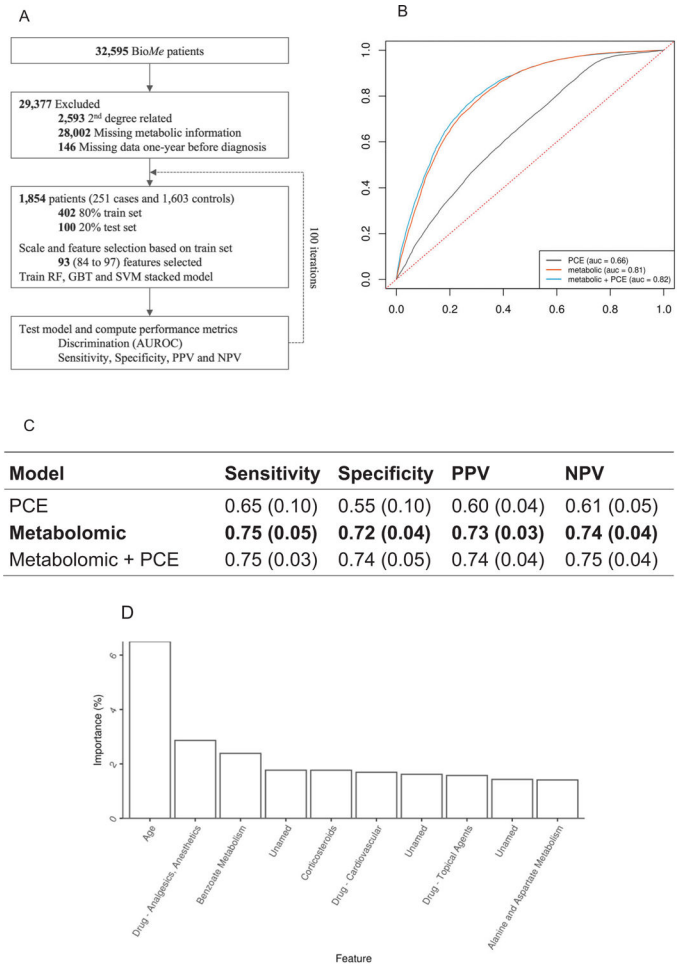

Fig. 1.

Study design and evaluation of metabolomic model for CAD prediction. A) Diagram describing filtering criteria, sampling, pre-processing, training and evaluation performed to generate a CAD prediction model using metabolomic information. RF corresponds to random forest; GBT corresponds to gradient boosted trees; SVM corresponds to support vector machines with polynomial kernel; AUROC corresponds to area under the receiver operator characteristic curve; PPV corresponds to positive predictive value and NPV corresponds to negative predictive value. B) Receiver operator characteristic curves. Y and X axis correspond to averaged true positive and false negative rates respectively across 100 iterations. The averaged area under the curve is indicated for every model. PCE corresponds to pooled cohort equations and auc corresponds to area under the curve. C) Per-class performance metrics. PCE corresponds to pooled cohort equations; PPV corresponds to positive predictive value and NPV corresponds to negative predictive value. D) Bar-plot showing feature importance on the metabolomic model. Feature importance is calculated as the normalized contribution of each of the three models and averaged across 100 iterations. Metabolomic signatures not be mapped to a biological process are labelled “Unamed”.