Towards Precise Predictive Modelling of Coronary Artery Disease Elaborating on Omics Data

Eleni I. Georga, *Student Member, IEEE*, Nikolaos S. Tachos, Antonis I. Sakellarios, *Member, IEEE*, Themis P. Exarchos, *Member, IEEE*, Gualtiero Pelosi, Oberdan Parodi, Lampros K. Michalis, and Dimitrios I. Fotiadis, *Senior Member, IEEE*

Abstract— **This study aims at developing a patient-specific model for coronary artery disease (CAD) risk stratification based on machine learning modelling of molecular, cellular, inflammatory and omics data.**

I. Introduction

Predicting the risk of coronary artery disease (CAD) constitutes a widely-studied problem from the perspective of statistical modelling. In spite of the reported good discrimination ability of parametric linear regression models, a recent systematic review demonstrated the paucity of external validation and head-to-head comparisons, the poor reporting of their technical characteristics as well as the variability in outcome variables, predictors and prediction horizons, which limits their applicability in evidence-based decision making in healthcare [1]. Precision medicine suggests dynamic individualized nonlinear predictive modelling approaches not being hypotheses-driven [2, 3].

II. CAD Risk Stratification

CAD risk stratification is formulated as a binary classification problem on the basis of a confined set of features (Table I), with a ≥50% diameter stenosis in at least one main coronary artery vessel, as assessed by CTCA, characterizing patients with mild to severe CAD. Three machine learning algorithms, ranging from parametric (i.e. feed-forward neural network) to non-parametric kernel-based ones (i.e. support vector machine) and ensemble models (i.e. random forest), have been examined. The discriminative capacity of the currently available data categories is evaluated by (i) a knowledge-based approach consisting in the a priori definition of 3 input cases (C1: Demographics, Risk Factors; C2: Demographics, Risk Factors, Symptoms; C3: Demographics, Risk Factors, Symptoms, Molecular Systemic Variables), and (ii) feature ranking according to the InfoGain criterion (C4). Table II reports classification results on 101 patients (No CAD: *n*=25, Age: 58.36±7.45; Mild to Severe CAD: *n*=76, Age: 63.61±7.43) by 10-fold cross-validation. The gradual improvement of accuracy with the enhancement of the input space is apparent, with proper customization of the input by feature ranking better balancing the sensitivity to specificity ratio. SVM outperforms MLP and RF resulting in an overall accuracy 85.1% and a nearly perfect sensitivity (98.7%), whereas specificity remains low (44.0%), presumably due to the class imbalance in the dataset. CAD risk stratification model refinement is ongoing by: (i) integrating new knowledge coming from big data sources (i.e lipid profile, exome and mRNA sequencing, exposome, inflammatory and monocyte markers), and (ii) selecting an effective modelling scheme advancing both the precision and interpretability of the results.

TABLE I. Dataset Description

Category	**Features**
Demographics	Age, Gender
Risk Factors	Family History of CAD, Hypertension, Diabetes, Dyslipidaemia, Smoking, Obesity, Metabolic Syndrome
Molecular Systemic Variables	Alanine Aminotransferase, Alkaline Phosphatase , Aspartate Aminotransferase, Creatinine, Gamma-Glutamyl Transferase, Glucose, HDL, High-Sensitivity C-Reactive Protein, Interleukin-6, LDL, Leptin, Total Cholesterol, Triglycerides, Uric Acid
Symptoms	Typical Angina, Atypical Angina, Non Angina Chest Pain, Other Symptoms, No Symptoms

TABLE II. Classification Performance

	MLP			**SVM**			**RF**		
	Acc.	*Se.*	*Sp.*	*Acc.*	*Se.*	*Sp.*	*Acc.*	*Se.*	*Sp.*
C1	66.3	78.9	28.0	77.2	97.4	16.0	73.3	85.5	36.0
C2	70.3	81.6	36.0	81.2	94.7	40.0	75.2	88.2	36.0
C3	74.3	84.2	44.0	84.2	97.4	44.0	77.2	97.4	16.0
C4	78.2	90.8	40.0	85.1	98.7	44.0	81.2	92.1	48.0

Acc. Accuracy, Se: Sensitivity, Sp: Specificity

References

[1] J. A. Damen, L. Hooft, E. Schuit, T. P. Debray, G. S. Collins, I. Tzoulaki, *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," *BMJ*, vol. 353, p. i2416, May 16 2016.

[2] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *European Heart Journal*, vol. 38, pp. 1805-1814, 2017.

[3] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nat Rev Cardiol*, vol. 13, pp. 350-9, Jun 2016.

*This work is funded by the European Commission: Project SMARTOOL, “Simulation Modeling of coronary ARTery disease: a tool for clinical decision support — SMARTool” GA number: 689068).

E. I. Georga, T. P. Exarchos, and D. I. Fotiadis are with the Unit of Medical Technology and Intelligent Information Systems, Materials Science and Engineering Department, University of Ioannina, Ioannina, GR 45110, Greece (corresponding author to provide phone: +301651009006; fax: +302651008889; e-mail: fotiadis@cc.uoi.gr).

N. S. Tachos, A. I. Sakellarios and D. I. Fotiadis are with the Institute of Molecular Biology and Biotechnology, Biomedical Research Department, FORTH, Ioannina, GR 45110, Greece.

G. Pelosi and O. Parodi are with the Institute of Clinical Physiology, National Research Council, Pisa IT 56124, Italy.

L. K. Michalis is with the Department of Cardiology, Medical School, University of Ioannina, Ioannina, GR 45110, Greece.