An interpretable data-driven approach with application to non-exercise based cardiorespiratory fitness stratification

D. Mendes[1], J. Henriques[1‡], P. Carvalho[1], T. Rocha[1,7], S. Paredes[1,7], R. Cabiddu[2], R. Trimer[2], R. Mendes[2], A. Borghi-Silva[2], L. Kaminsky[3], E. Ashley[4], R. Arena[5], J. Myers[6]

[1]CISUC, Department of Informatics Engineering, University of Coimbra, Portugal. {dianasxm, jh, carvalho}@dei.uc.pt, {teresa, sparedes}@isec.pt.

[2]Cardiopulmonary Physical Therapy Laboratory, Department of Physical Therapy, Federal University of Sao Carlos, Sao Carlos - Sao Paulo, Brazil, retrimer@hotmail.com, {ramona.cabiddu, mendesrg, audreyborghi}@gmail.com,

[3]Fisher Institute of Health and Well-Being and Clinical Exercise Physiology Laboratory, Ball State University, Muncie, IN, United States - kaminskyla@bsu.edu.

[4]Division of Cardiovascular Medicine, VA Palo Alto Healthcare System, Palo Alto, CA, and Stanford University, Stanford, CA, United States - euan@stanford.edu.

[5]Department of Physical Therapy and Integrative Physiology Laboratory, College of Applied Science, University of Illinois, Chicago, IL, United States - rarena70@gmail.com.

[6]Department of Cardiovascular Medicine, Stanford University, Palo Alto, CA, United States - drj993@aol.com.

[7] Department of Informatics Engineering and Systems, Polytechnic Institute of Coimbra / ISEC, Coimbra, Portugal

‡ Correspondent author: {jh@dei.uc.pt}

Abstract

Objectives: The continued exploration of clinically relevant predictive models continues to be an important pursuit. The aim of the current study was to develop a data-driven model, based on computational intelligence techniques, to predict the maximum oxygen consumption as a measure to be used in cardiorespiratory fitness stratification. While the maximum oxygen consumption is a direct mark of the cardiorespiratory fitness, several studies have also confirmed it as a powerful predictor of risk for adverse outcomes, such as hypertension, obesity, and diabetes. Therefore, the existence of simple and accurate models, establishing an alternative to standard cardiopulmonary exercise tests, with the potential to be employed in the stratification of the general population in daily clinical practice, would be of major importance.

Methods: The primary hypothesis to be explored in this work is that individuals with similar characteristics present similar cardiorespiratory fitness levels. Therefore, this work addresses the development of data-driven stratification models able to learn distinct groups (classes) of subjects assessing the similarity between characterizing variables. Moreover, the stratification scheme should permit the definition of interpretable models that characterize the distinct subjects, aligned with the clinical evidence. While increasing clinician confidence, the extracted information can be used to complement clinical evidence.

Results and conclusions: The models' performance was evaluated using the "FRIEND - Fitness Registry and the Importance of Exercise: The National Data Base". A subset of N=10887 healthy individuals was employed being the proposed cluster approach compared with the traditional Wasserman/Hansen equations. Accuracy (geometric mean) results show the superiority of the proposed approach in the prediction of maximum oxygen consumption. While presenting higher accuracy results, clustering approach can be seen as a valid alternative to traditional models mainly due to their modularity, robustness, capacity do deal with non-balance data and to extract meaningful information from available clinical datasets.