Identifying Ventricular Arrhythmia Cases and their Predictors by Applying Machine Learning Methods to Electronic Health Records (EHR) of Hypertrophic Cardiomyopathy (HCM) Patients*

Moumita Bhattacharya[1], Dai-Yin Lu, MD[2], Prasanth Lingamaneni, MD[2], Shibani Kudchadkar, MD[2], Gabriela Villareal, MD[2], Sanjay Sivalokanathan, MD[2], Pam Corona Villalobos, MD[3], Stefan Zimmerman, MD[3], Theodore P. Abrahram, MD[2], M. Roselle Abraham, MD[2] and Hagit Shatkay, PhD[1,2]

[1]Computational Biomedicine Lab, Computer Sciences, University of Delaware, Newark, DE, USA; [2] Hypertrophic Cardiomyopathy Center of Excellence, Division of Cardiology/Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA; [3]Department of Radiology, Johns Hopkins School of Medicine, Baltimore, MD, USA

Introduction: Ventricular arrhythmias (VA) are a major cause of death in hypertrophic cardiomyopathy (HCM) patients. We develop a machine learning method to effectively identify VA in HCM patients using a small set of clinical variables.

Methods: We scanned the EHR of *788* HCM patients, who underwent detailed clinical phenotyping, for sustained ventricular tachycardia/fibrillation. Patients with VA (*61*) were tagged as *Arrhythmia cases* and the remaining (*727)* as *non-Arrhythmia*. To identify the most informative variables for separating arrhythmia from non-arrhythmia we used the *2-sample t-test* and *power analysis*. Patient records were reduced to include only these variables.

Notably, the dataset is highly imbalanced, resulting in poor performance of conventional classifiers. While imbalance is often addressed by either over- or under-sampling of one of the classes, we apply a combination of both. We trained and tested multiple classifiers (including random forest and logistic regression), under this sampling strategy, showing effective classification.

Results: Of the 160 measured variables, 21 were informative for identifying patients with VA (Table 1). The logistic regression classifier, trained based on these variables and corrected for data imbalance, is most effective in separating *Arrhythmia* from *non-Arrhythmia* patients (~0.73 *sensitivity*, ~0.80 *specificity*, ~0.78 *C-index*). The performance is significantly higher than that of a recent method by O'Mahony et al., (*C-index* ~0.69). Our study also reveals several predictive variables (e.g. *myocardial strain measurements*) that have hitherto not been associated with VA prediction (see Table 1).

Conclusions: This is the first application of machine learning for identifying VA using clinical variables. A small subset of variables, many of which not used before, proved effective for prediction. Prospective testing is needed to demonstrate improved risk prediction of VA using this set of variables.

* This work was funded in part by NSF IIS EAGER grant #1650851 and an award from the John Taylor Babbitt foundation.

Table 1: Variables identified through feature selection (*21* of the original *160 variables*), as *highly informative* of Ventricular Arrhythmia (*VA*), shown in descending order of their respective information value.
Variables that have not been associated with VA prediction before are shown in boldface in cells highlighted in grey. '+' indicates that the variable has a *higher* value in patients with *VA*, compared to *Non-Arrhythmia patients*, while '−' indicates a *lower* value of the variable in patients with *VA*. For *categorical variables*, (e.g. *History of Syncope, HCM Type*), the value most highly associated with arrhythmia is also shown.

Variables highly informative of Ventricular Arrhythmia
Peak pressure gradient at left ventricular outflow tract (*LVOT*) *at peak stress* ($LVOTG_{Stress}$) (−)
History of Syncope (Presence +)
HCM Type (Non-obstructive +)
Systolic blood pressure before treadmill exercise (−)
Global longitudinal early diastolic strain rate (−)
Maximal thickness of interventricular septum (+)
Global longitudinal systolic strain rate (−)
Global longitudinal systolic strain, % (−)
Exercise Time (−)
Syncope (Presence +)
Ejection fraction (%) of left ventricle (−)
IVS max to PW ratio (IVS/PW), where PW is maximal thickness of posterior wall of left ventricle (+)
Diastolic blood pressure before treadmill exercise (+)
Metabolic equivalents (−)
Presence/absence of inducible VT by EP study during follow-up (Presence +)
Body Mass Index (−)
Age (−)
Peak pressure gradient at the left ventricular outflow tract, at rest ($LVOTG_{Rest}$) (+)
Family History of HCM (Presence +)
Family History of Sudden Cardiac Death (Presence +)
Ratio of early diastolic filling velocity to late diastolic filling velocity (−)