ARIC Manuscript Proposal # 3123

PC Reviewed: 2/13/2018  Status: _____  Priority: 2
SC Reviewed: _________  Status: _____  Priority: _____

1.a. Full Title: Machine learning based phenotyping in heart failure

b. Abbreviated Title (Length 26 characters): Phenotyping in heart failure

2. Writing Group:
   Writing group members: Sergio Sanchez Martinez, Maja Cikes, Amil Shah, Brian Claggett, Susan Cheng, [others welcome], Bart Bijnens, Scott D. Solomon

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. __SSM__ [please confirm with your initials electronically or in writing]

First author: Sergio Sanchez Martinez
Address: Universitat Pompeu Fabra
         Department of Information and Communication Technologies
         Roc Boronat, 138.
         Tanger Building, office 55.119
         08018, Barcelona

Phone: (+34) 636200177  Fax: (+34) 93 542 25 17
E-mail: sersanmarsergio@gmail.com

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).
Name: Scott D. Solomon
Address: Brigham and Women’s Hospital
         Cardiovascular Division
         75 Francis Street
         Boston, MA 02115

Phone: 857-307-1960  Fax: 857-307-1944
E-mail: ssolomon@rics.bwh.harvard.edu
3. **Timeline:** Analysis will begin following proposal approval with the aim of completing analysis and a manuscript within 6 months of data becoming available.

4. **Rationale:**
   
   **Purpose/Aims:**
   We aim to further define patient phenotypes and identify their natural history by harnessing the power of state-of-the-art supervised as well as unsupervised machine learning techniques, utilizing the rich echocardiographic data available from the ARIC database.

   **Background:**
   Distinct phenotypes and etiologies are present in various subgroups of HF patients, but are not well defined, particularly in HF with preserved ejection fraction (HFpEF). However, this specific syndrome is particularly heterogeneous and associated with comorbidities [1, 2, 3].

   Traditional methods to define phenotypes within groups of individuals with and at risk for HF rely on elucidation of individual phenotypic subgroups that can be arbitrarily defined and focus on unidimensional subgroups (i.e., presence or absence of individual comorbidities, risk factors or measured assessments).

   While assessment of cardiac structure and function using current echocardiographic analysis tools can identify subgroups of individuals who are at increased risk for developing HF [4, 5], standard approaches ascribe risk to a limited amount of individual measurements in a unidimensional fashion. Machine learning techniques offer the advantage of enabling the use of the full acquired data (either the images themselves or the time-varying traces, e.g. strains, blood velocities etc. - extracted from them) without assumptions on which single measurements are most relevant for the patient population under study. Additionally, the richness of the data provides enough information to enable identification of clusters of patients with similar (but not identical) properties without prior assumptions on clinical/diagnostic labels. This highlights the properties of the images/data, previously unexplored, thus acting as novel hypothesis-generating assessments [6]. We have previously demonstrated feasibility utilizing previously segmented strain contours from echocardiographic images to identify distinct subgroups based on patterns of cardiac deformation.

5. **Main Hypothesis/Study Questions:**
   
   Our central hypothesis is that machine learning may provide better ways to identify phenotypes using multidimensional data and relate these phenotypes to the development of heart failure, and thus identify the pertinent variables, without *a priori* (mostly historical) assumptions, that increase risk for heart failure.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary
of data analysis, and any anticipated methodologic limitations or challenges if present).

Contemporary machine learning approaches allow for rearranging large amounts of data for either classification (mostly supervised) or exploration (unsupervised) [7, 8]. The approach we have developed is based on a non-linear dimensionality reduction scheme after which clustering is performed in the resulting low-dimensional space in order to identify phenogroups of similar patients. Additionally, when follow-up data is available for each of the agnostically determined phenogroups (one of which will be the ‘normal’ individual, corresponding to the average of the cluster representing the majority of event-risk factor-free individuals), these phenogroups can be distinguished by outcomes and decomposed into the clinical and echocardiographic characteristics that are more abundant within each phenogroup.

Analyzing strain data derived from echocardiograms of patients from a large CRT trial, we have provided proof-of-concept that machine-learning based approaches can identify clinically distinct patterns from deformation imaging in a phenotypically heterogeneous cohort, thus potentially aiding in better recognition of disease-specific patterns that might benefit from specific therapies. However, in order to investigate a broader phenotypical population, such as heart failure or pre-heart failure, a large population (> 5000 individuals) will be necessary to obtain a stable low-dimensional space that wouldn’t significantly change when a few individuals are added.

Big data based machine learning has the additional advantage that a large number of heterogeneous variables can be used to create the low-dimensional space. This is achieved by ordering patients according to similarity of data entities (e.g. strain traces, Doppler patterns, demographics, risk factor data…) and subsequently defining the properties of the data that can summarize the difference in these entities along more and more dissimilar patients in the most efficient way (the most commonly used linear variant of dimensionality reduction is Principle Component Analysis [9]).

The proposed learning approach, as well as the non-linearity of the method [10], will identify the variables contributing most to the best separation of the phenogroups. Therefore, using as many imaging variables as would be available would provide the best results. Strategies to deal with missing variables in some of the individuals are available. The result of the learning and clustering phase is the identification of phenogroups and how these phenogroups relate to the echocardiographic images/parameters on which they are based.

Each individual used during the learning phase will thus be positioned within this space/groups. Additionally, the ‘representative’ (=average) of each group will be available for intuitive representation of the typical clinical/echocardiographic features of each cluster. When follow-up data are available, changes in the positioning within the clusters can be assessed, thus extracting individual trajectories over time. The association of an outcome variable (death, hospitalization, transition to HF, change in echo parameters…) with the positioning allows for evaluation of the temporal evolution of the individual, which can be subsequently used to predict the prognosis of new individuals that are positioned within the learned clusters.
Analysis Methods:
The proposed methodology is based on the one developed and described in a recent publication by Sanchez-Martinez et al [11]. Given the vast amount of information present in echocardiographic examinations, since the dimensionality of each descriptor is determined by the number of entries that it has (e.g. for myocardial strain traces, each entry is the instantaneous myocardial strain at a certain time-point), the first step involves dimensionality reduction towards a space that highlights the main characteristics of the input data. Prior to this, the cardiac phases of the echocardiographic patterns (e.g., strain traces or Doppler flows) need to be time-aligned. Then, affinities between each pair of subjects are computed for each descriptor. Based on these affinities, the learning process maps the input data to a space of lower dimensionality that positions subjects according to their similarity. This is allowed by the multiple kernel learning (MKL) algorithm we have developed, which handles different descriptors by automatically weighing their relative importance to the final result. In practice, the number of dimensions of the achieved space equals the number of evaluated subjects minus 1. However, the most salient characteristics of the data are reflected in the first few dimensions. The low-dimensional space captures variations in the whole population and within each subgroup, and facilitates interpretation of the data.

Our algorithm is unsupervised, meaning that the machine learning is not conditioned by prior clinical diagnostic labels which may be inaccurate, or outcomes, but instead is equivalent to an independent blinded interpretation.

The low-dimensional space provides distances between subjects that indicate how similar they are, but no labels are attributed. We investigate if this space could be used to automatically assign subjects to a given class, which is achieved through agglomerative hierarchical clustering over the subject coordinates. The clustering seeks to identify different characteristic patterns within the analyzed subjects in an unsupervised way.

However, the interpretation of the groupings identified by clustering on the low-dimensional space remains physiologically abstract. Thus, we then evaluate whether the learned clusters are clinically meaningful by comparing relevant clinical parameters and by studying the variability of input descriptors associated to them. This variability is reconstructed using advanced regression techniques and can be interpreted in light of the pathophysiologic characteristics of heart failure. Our machine learning approach not only assigns class-labels to individuals, corresponding to e.g. a diagnosis/etiology or predicted outcome, but additionally provides a distance measure with regards to other groups, thus giving a ‘severity’ measure for each individual.

Last, we describe outcome data among clusters using Kaplan–Meier estimates for heart failure or death. Our learning approach can also be used to predict outcome (e.g., heart failure or death) at an individual level by following a nearest neighbor approach. To this end, a subject from a validation set would be projected to the low-dimensional space learned by our ML model, and based on the outcome data of the most similar subjects in the training set, a prediction for that subject would be made.

7.a. Will the data be used for non-CVD analysis in this manuscript?
   ___ Yes  ___ No
b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used?
   _____ Yes  X No

(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?
   _____ Yes  X No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = “No use/storage DNA”?
   _____ Yes  X No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.csc.unc.edu/ARIC/search.php
   X  Yes  _______ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

   Amil M Shah, Kunihiro Matsushita, Dalane Kitzman, Ervin Fox, Suma Konety, Scott D. Solomon; Others welcome. The relationship between concentric remodeling and left ventricular function – A preliminary analysis from the ARIC study.

   Tor Biering-Sørensen, Amil Shah, Kotaro Nochioka, Maja Cikes, Brian Claggett, Scott D. Solomon, others welcome. Regional left ventricular deformation obtained by speckle tracking echocardiography in different risk groups and in different spectra of heart failure.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data?
   X  Yes  ___ No

11.b. If yes, is the proposal
   _____ A. primarily the result of an ancillary study (list number* 2017.05, approved ancillary study: “Heart Failure-tailored decision support: from Precision to Personalized Medicine” The Grant was, however, not funded, but we are continuing with research in the field)
B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________  __________  __________

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is your responsibility to upload manuscripts to PUBMED Central whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.

References


