Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction

Sergio Sanchez-Martinez, Nicolas Duchateau, Tamas Erdei, Gabor Kunszt, Svend Aakhus, Anna Degiovanni, Paolo Marino, Erberto Carluccio, Gemma Piella, Alan Fraser, et al.

CIRCCVIM/2017/007138 R2

Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction

Running title: *Sanchez-Martinez et al.; Machine learning to characterize HFPEF*

Sergio Sanchez-Martinez[a], MSc.; Nicolas Duchateau[b], PhD; Tamas Erdei[c], MD, PhD; Gabor Kunszt[d], MD; Svend Aakhus[d,e,f], MD, PhD; Anna Degiovanni[g], MD; Paolo Marino[g], MD; Erberto Carluccio[h], MD; Gemma Piella[a], PhD; Alan G. Fraser[c], MD*; Bart H. Bijnens[a,i], PhD*

[a] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain

[b] Inria Asclepios research project, Sophia Antipolis, France

[c] Wales Heart Research Institute, Cardiff University, United Kingdom

[d] Department of Cardiology, Oslo University Hospital, Oslo, Norway

[e] Department of Circulation and Imaging, Faculty of Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

[f] Clinic of Cardiology, St. Olav Hospital, Trondheim, Norway

[g] Department of Cardiology, University of Eastern Piedmont, Novara, Italy

[h] Division of Cardiology, University Hospital "S.Maria della Misericordia", Perugia, Italy

[i] ICREA, Barcelona, Spain

Address for correspondence:
Sergio Sánchez Martínez
DTIC, Universitat Pompeu Fabra (office 55.119)
Roc Boronat 138, E08018 Barcelona, Spain
Tel.: (+34) 935421346
Fax: (+34) 935422517
E-mail: sergio.sanchezm@upf.edu

* Drs. Fraser and Bijnens contributed equally to this work

CIRCCVIM/2017/007138 R2

Abstract

Background – Current diagnosis of heart failure with preserved ejection fraction (HFPEF) is suboptimal. We tested the hypothesis that comprehensive machine learning (ML) of left ventricular (LV) function at rest and exercise objectively captures differences between HFPEF and healthy subjects.

Methods and results – 156 subjects aged>60 years (72 HFPEF+33 healthy for the initial analyses; 24 hypertensive+27 breathless for independent evaluation) underwent stress echocardiography, in the MEDIA-study. LV long-axis myocardial velocity patterns were analyzed using an unsupervised ML algorithm that orders subjects according to their similarity, allowing exploration of the main trends in velocity patterns. ML identified a continuum from health to disease, including a transition zone associated to an uncertain diagnosis. Clinical validation was performed: (i) to characterize the main trends in the patterns for each zone, which corresponded to known characteristics and new features of HFPEF; the ML-diagnostic zones differed for age, body mass index, 6-minute walk distance, B-type natriuretic peptide, and LV mass index ($p<0.05$). (ii) to evaluate the consistency of the proposed groupings against diagnosis by current clinical criteria; correlation with diagnosis was good (Kappa, 72.6%; 95% confidence interval, 58.1–87.0); ML identified 6% of healthy controls as HFPEF. Blinded reinterpretation of imaging from subjects with discordant clinical and ML diagnoses revealed abnormalities not included in diagnostic criteria. The algorithm was applied independently to another 51 subjects, classifying 33% of hypertensive and 67% of breathless controls as mild-HFPEF.

Conclusions – The analysis of LV long-axis function on exercise by interpretable ML may improve the diagnosis and understanding of HFPEF.

Key words: exercise echocardiography; heart failure with preserved ejection fraction; myocardial velocity; machine learning; non-invasive diagnostics heart failure.

Introduction

Heart failure with preserved ejection fraction (HFPEF) results from multiple pathophysiologic processes but diagnostic criteria remain general, including dyspnea and fluid overload, normal left ventricular (LV) ejection fraction (EF), elevated natriuretic peptides, and evidence of heart failure or diastolic dysfunction[1].

EF may not reveal LV long-axis systolic dysfunction[2] and resting diastolic function can be normal[3]. Diagnosis relies on echocardiography at rest while abnormalities may appear only during exercise[4]. In case of uncertainty, the diagnosis may be confirmed by a stress test or elevated LV filling pressure.

Negative results of trials investigating HFPEF therapies may be due to the limitations of current diagnostic criteria[5]. Alternative approaches combining clinical and imaging indexes[6,7] may not incorporate enough measurements to capture the complexity of HFPEF. Clinical studies tend to measure what we know and recognize, using scalar indexes, while interrogating patterns of cardiac function may be more informative. In that context machine learning (ML), which allows all the data to be considered, may be insightful. Supervised ML, a configuration that is trained using labels (e.g., clinical diagnosis), is becoming successful for classification[8,9]. In patients with suspected heart failure ML should be unsupervised – meaning that it is performed independently of diagnostic labels – so that it is not biased by possibly erroneous diagnoses.

Invasive measurements in subjects with HFPEF have shown increased filling pressures, exercise-induced pulmonary hypertension and blunted functional reserve[10], but their use is limited in clinical practice, giving echocardiography a central role in the diagnosis of HFPEF. Exercise echocardiography has been advocated for the early diagnosis of HFPEF[3,11,12], to stratify risk[13] and to estimate prognosis[14]. It can differentiate between causes of decreased functional reserve, such as inability to enhance myocardial relaxation, increased chamber stiffness with elevated LV filling pressure, and exercise-induced pulmonary hypertension[2].

Previous studies confirmed that quantifying long-axis responses to stress can detect myocardial ischemia and diagnose coronary artery disease[15] and that analysis of regional long-axis function is informative about myocardial mechanics[16].

We hypothesized that unsupervised ML using basal myocardial long-axis velocity patterns at rest and exercise would discriminate between healthy and HFPEF subjects with impaired functional reserve, and would identify new descriptors that better characterize the HFPEF syndrome.

Methods

Study population

We collected data from four centers of the MEDIA-study (MEtabolic Road to DIAstolic Heart Failure): University Hospital of Wales (UK), Scuola di Medicina of Eastern Piedmont University (Italy), Università degli Studi di Perugia (Italy), and Oslo University Hospital (Norway). These data will not yet be available to other researchers for reproducibility purposes until the publication plan of the MEDIA-study has concluded.

156 subjects aged ≥60 years were recruited into 4 subgroups: (i) patients with HFPEF; (ii) breathless patients without HFPEF; (iii) asymptomatic hypertensive subjects; and (iv) healthy controls. HFPEF was diagnosed according to the 2007 recommendations from the European Society of Cardiology (ESC), namely symptoms or signs of heart failure, LVEF >50%, and a non-dilated LV (end-diastolic volume index <97 mL/m^2) with evidence of abnormal LV relaxation, filling, diastolic distensibility, or diastolic stiffness, and/or an elevated N-terminal pro-brain natriuretic peptide (NTpro-BNP) concentration, and/or left atrial enlargement, and/or atrial fibrillation[17]. Patients with dyspnea on exertion not meeting the previous criteria were recruited as “breathless” controls. Asymptomatic volunteers aged >60 years without diabetes or any cardiovascular disease, were recruited as healthy controls. If their blood pressure (BP) was mildly elevated (systolic BP>140 mmHg and/or diastolic BP>90 mmHg) they were categorized

as hypertensive controls. Exclusion criteria for all groups included any severe respiratory cause of dyspnea such as asthma or chronic obstructive pulmonary disease; acute or previous myocardial infarction or known coronary artery disease awaiting revascularization; and cerebrovascular disease or stroke within the previous 3 months.

Ethical approval was given by the Ethics Committee of each institution, and each subject gave written informed consent.

Echocardiography

All subjects underwent echocardiographic studies at rest and during exercise using a semi-supine bicycle with a ramped protocol[11]. If the subject developed symptoms or once she/he reached a heart rate of 100/min, the workload was held constant for 3 minutes while imaging was performed during submaximal exercise. All centers used a Vivid E9 echocardiographic system with an M4S transducer (GE Healthcare, Milwaukee, WI).

Three-beat loops of apical 4-chamber tissue Doppler images were acquired at a sampling rate of 180±34 Hz and analyzed using commercial software (EchoPAC, v.113, GE Healthcare). Velocity traces were extracted from LV basal septal and lateral segments, using a sample size of 1×10 mm placed 10 mm above the mitral annulus in systole, to avoid capturing ring motion. Manual or automatic (speckle-) tracking of the sampling points introduced additional variability without significant changes on the traces ; therefore we avoided tracking not to compromise reproducibility[18]. One beat was analyzed for every subject in the study.

Temporal normalization

To allow quantitative comparisons between traces with different heart rates and timing of cardiac phases, they were temporally aligned, using the timeline of the most typical subject (closest to the average among controls) as reference. Events were defined from valve flows for each subject and during each stage of exercise: mitral valve closure, aortic valve opening, aortic valve closure, mitral valve opening, and onset of atrial contraction. A two-step process was

used: (1) phase-wise warping, to ensure temporal coincidence of cardiac events; and (2) resampling to the reference, to ensure equal numbers of sampling points for the analyses[19].

Machine learning

The main steps of our algorithm are shown in Figure 1. The input consisted of 22 descriptors (Figure 2). Twenty corresponded to the 5 phases of 4 velocity traces (septal and lateral at rest and submaximal exercise) – isovolumic contraction, systolic ejection, isovolumic relaxation, early diastole including diastasis, and late diastole (atrial contraction). We reported previously that diagnostic information is captured not only by the amplitude of velocity, but also by the relative changes in duration of the cardiac phases[18]. Thus, we added 2 extra descriptors that consist of the timings of each subject's physiologic events as compared to the reference, one for the normalization at rest and the other at exercise.

The population analyzed during learning consisted of 105 subjects: 33 healthy volunteers, and 72 HFPEF patients[17]. The ML model was then evaluated independently in two additional cohorts: 27 breathless and 24 hypertensive subjects.

Dimensionality reduction: The dimensionality of velocity patterns equals the number of instantaneous acquisitions that they have. Our input was high-dimensional – for example, 22 descriptors reaching up to 300 dimensions in the case of the early diastolic phase.

The learning process computed a dimensionality-reduced space that preserved the similarities between each pair of subjects calculated for each descriptor (Figure 1, step #1 and step #2[18,20]). Our dimensionality reduction formulation was unsupervised, i.e., blinded to diagnostic labels since they might be inaccurate. Specifically, we used unsupervised multiple kernel learning, a previously validated ML algorithm[18], which handles heterogeneous descriptors and reduces their complexity into a low-dimensional space. The number of dimensions of the achieved space equals the number of evaluated subjects minus 1; 104 in this study. Nonetheless,

we only considered the first few dimensions, which generally capture the most salient characteristics of the data[18] (step #3), and facilitate interpretation of the trends in the population.

Clustering: The low-dimensional space preserves similarities between subjects without attributing (diagnostic) labels. We harnessed its potential to agnostically group subjects in two classes using agglomerative hierarchical clustering[21] (step #4), targeted to capture the healthy and diseased characteristic patterns (of cardiac motion) within the population. In practice, clustering was performed assessing dissimilarity and linkage via the Euclidean distance and Ward's criterion (to minimize the intra-cluster variance), respectively.

Clinical validation

Variability analysis: After the learning process, we assessed the clinical relevance of the clusters by comparing diagnostic parameters among them (step #5) and by studying their trends in velocity patterns (step #6). These trends were described among clusters using principal component analysis (PCA), to find their main modes of variation, coupled with regression techniques[18], which computed the variability of velocity patterns explained by these modes. Note that the PCA was not intended to further reduce the dimensionality of the data, but just as a tool to describe clusters.

Clusters versus clinical labels – uncertainty in the diagnosis: Based on the prevalence of clinical labels within the two clusters we identified which represented the "healthy" and which the "HFPEF" characterizations. Next, we quantified membership probabilities for each subject based on their Mahalanobis distance to the barycenter of each cluster. Thus, we defined regions in the low-dimensional space corresponding to "healthy" and "HFPEF", as well as an intermediate "transition zone", whose cut-points were selected to maximize the discordant cases (whose probability by ML differed from clinical diagnosis) while minimizing the concordant cases (step #7; Supplementary Figure S1). We did not expect full agreement between ML and clinical diagnosis, as our objective was to find new (data-driven) groupings that could be more

instructive than the possibly suboptimal consensus recommendations. Blinded re-analysis of the discordant diagnosis cases was performed. The details are provided in the Data Supplement.

Independent testing on separate patient groups

After learning from the 105 healthy and HFPEF subjects, the diagnostic algorithm was evaluated independently in two additional cohorts: 27 breathless and 24 hypertensive patients, which were mapped to the healthy, HFPEF or transition regions (Figure 1, step #8).

Statistical analysis

Categorical variables are expressed as counts and percentages, and group differences were assessed using the chi-square test. Continuous variables that were found to be non-normally distributed are presented as median with 25th to 75th percentiles; inter-class differences were calculated by the non-parametric Kruskal-Wallis test. A p-value of less than 0.05 was considered statistically significant. Agreement between ML and clinical labels was expressed by the Kappa statistic. The ML algorithm and the statistical analyses were implemented using MATLAB (R2016b, The MathWorks Inc., Natick, MA, 2016).

Results

By definition, HFPEF subjects had higher NT pro-BNP, E/e' ratio, LV mass index (LVMI), and left atrial volume index (LAVI), than the healthy controls (table 1). On average, they were 5.1 years older, had higher body mass index (BMI) and shorter 6-minute walk test (6MWT) distance. The median heart rates during submaximal exercise were 102 (100–106) min^{-1} in healthy subjects compared with 100 (90–107) min^{-1} in HFPEF (p=0.042). There were no major differences between subjects from different participating centers (Supplementary table 1).

Machine learning

The first 10 dimensions of the low-dimensional space were considered for clustering, as they encode the highest variability in the pattern data.

Subjects in cluster 2 were 6% older, had higher BMI (by 13%), NTpro-BNP (by 85%), and LVMI (by 28%), and their 6MWT distance was 31% shorter than subjects in cluster 1 (table 1; all $p<0.05$). The E/e' ratio was higher in cluster 2 at rest (+9%, $p=0.028$) but similar during submaximal exercise (+11%, $p=0.446$).

Based on these comparisons and the prevalence of diagnostic labels within the clusters, we considered clusters 1 and 2 as *healthy* and *diseased* clusters, respectively. There were no significant differences between the *diseased* cluster identified by ML (n=79) and the HFPEF group defined by applying clinical criteria (n=72) in any of the standard variables (table 1). The *healthy* cluster (n=26) and the clinically-defined healthy group (n=33) differed only in E/e' during exercise (14% higher in the *healthy* cluster, $p=0.048$).

Clinical validation

Variability of the clusters: The variability corresponding to the first two cluster modes is shown in Figure 3. The *diseased* cluster showed lower velocities, more fusion of early and late diastolic curves during exercise, higher variability in the onset of atrial contraction, and smaller increase in myocardial velocity corresponding to atrial contraction during exercise.

Figure 4 summarizes differences between clusters in clinically interpretable features up to the tenth cluster mode. This confirms that amplitudes of velocity were higher in the *healthy* cluster. Diastolic fusion was more pronounced in the *diseased* cluster, particularly in the septum during exercise − perhaps because of delay in the onset of diastolic filling (also shown by timing bars in Figure 3). The *diseased* cluster also showed more variability in systolic and diastolic duration (1st mode) and more frequent inter-atrial contraction delay (2nd and 5th modes).

Diagnostic relevance of the clusters: Moderate agreement was observed between the learned clusters and the diagnostic labels (Kappa, 72.6%; 95% confidence interval, 58.1–87.0); 22 out of 105 subjects were classified differently by ML (Figure 5, table 2). The Mahalanobis distance from each subject to the center of each cluster is depicted in Figure 6A; the greater the distance

to the opposite cluster, the higher the probability of correct diagnosis. For intermediate probabilities, we defined a transition zone between the clusters, denoting a high uncertainty in binary diagnosis (more details in Supplementary Figure S1). A blinded re-analysis of the discordant diagnosis cases is provided in the Data Supplement.

Independent evaluation in breathless and hypertensive subjects

Hypertensive and breathless controls were mapped to the low-dimensional space and their distances to the learned clusters were calculated (Figure 6B-6C).

All hypertensive subjects mapped to the transition zone (n=16; 67%) or the milder part of the HFPEF region (n= 8; 33%) (table 2), with their distance from the *healthy* cluster being moderately related to their resting systolic BP (Pearson coefficient r=0.51, p=0.07). Most breathless subjects mapped to the transition zone (n=8; 30%) or the milder part of the HFPEF region (n=18; 67%).

Discussion

Our study is the first to apply machine learning to analyze myocardial long-axis motion throughout the cardiac cycle and during exercise. We confirmed the hypothesis that this method can identify groups of subjects with different cardiac functional reserve, measured by echocardiography. We demonstrated that the diagnosis of HFPEF based on consensus recommendations may fail to identify some patients with a cardiac cause for their symptoms while also designating others as diseased when their response to exercise is healthy (see Data Supplement).

We used unsupervised learning because of doubts that diagnostic criteria, limited to resting cardiac assessment, can identify all subjects with the HFPEF syndrome. Dimensionality reduction and clustering blindly identified clinically distinct groups that share similarities with diagnostic recommendations[17], objectively quantified the difference from a control group, and

described a "transition zone" where standard criteria would have a lower diagnostic accuracy. This suggests that ML can offer an objective method for diagnosing heart failure.

We studied LV long-axis function because it is reduced in HFPEF patients[22] and because tissue Doppler imaging provides high temporal resolution and reproducible signals that can be easily extracted and post-processed. We selected patients with HFPEF and healthy controls, using consensus definitions, but studied them in a blinded fashion to develop the model. We enrolled two intermediate diseased groups − asymptomatic hypertensive subjects, and breathless patients who did not fulfil HFPEF diagnosis− to re-assess the learned model in independent populations. We did not use speckle tracking to quantify longitudinal strain, because strain is preload-dependent[23], and thus less appropriate than myocardial velocity or strain rate as an index of contractile function and reserve.

Advantages of machine learning

Pathophysiologic processes associated with HFPEF – such as systemic inflammation, LV hypertrophy, LV diastolic stiffness, and left atrial remodeling – may progress continuously from health to disease. Clinical measurements may be normally distributed, such that the definition of diagnostic cut-points becomes difficult or even arbitrary. Our unsupervised ML model is advantageous as it eschews categorical diagnoses, which might be biased, in favor of providing membership probabilities to diseased or healthy groups or a quantitative estimate of divergence from normality. It is therefore appropriate to discriminate between heterogeneous phenotypes that are currently lumped together within the HFPEF syndrome[6]. We used it to separate the subjects into two main groups (healthy and diseased) but larger numbers would allow clustering into more specific HFPEF phenotypes. Setting more clusters would allow machine learning to capture finer patterns, but at the risk of (over)fitting.

Two previous studies sought to classify patients with HFPEF, but their analyses were limited to sets of 11 and 67 scalar variables, without functional data during exercise[6,7]. We

analyzed patterns rather than scalar indexes and extracted their most salient characteristics by keeping the first 10 dimensions of the dimensionality-reduced space, discarding the rest to prevent overfitting.

Diagnostic recommendations rely heavily on LVEF and E/e' ratio but both are controversial[11,7,24]. We have demonstrated that characterizing subjects based on their complex patterns of myocardial motion at rest and during exercise would be more informative. Indeed, our analysis revealed undiscovered diagnostic features on the motion patterns. It could be argued that our variability analysis is equivalent to performing comparisons on instantaneous velocities independently, but that approach did not reveal clear differences between healthy, hypertensive and breathless subjects (Supplementary Figure S2).

Among ML techniques, deep-learning has captured most attention since it performs well in challenging tasks such as segmentation. It is now a mature method for extracting features that can be analyzed within a supervised model[25], but its "black box" nature hinders interpretation of the results. In contrast, our method remains clinically interpretable, since it gives insights into the meaning of the clusters through the variability analysis.

Pathophysiologic interpretation

The identified clusters were clinically relevant – most diagnostic parameters[17] differed between them. We complemented the learning with a physiologic interpretation of the pattern trends associated to the clusters. The *diseased* cluster showed lower systolic and diastolic amplitudes, indicating impairment of functional reserve; more fusion of early and late diastolic curves during exercise (at similar heart rates), which may come from increased late systolic wave reflections delaying early diastolic lengthening, or from an interaction between relaxation and compliance (or early and late diastolic filling); increased variability in the onset of atrial contraction (a' wave), which might be the result of diastolic and inter-atrial dyssynchrony, as

recently reported in HFPEF[26]; and a blunted response in atrial velocities (a' wave peak), failing to increase during exercise, suggestive of increased filling pressure. Some of these are not yet considered as diagnostic features of HFPEF, and so they merit further investigation.

Direct use of the learned clusters to allocate breathless patients into two distinct groups − with or without HFPEF − would be unrealistic due to the continuous transition from health to disease that we confirmed across the four studied groups[27]. This supports the view that current diagnostic criteria for HFPEF are suboptimal. We propose instead that automated diagnosis could be supported by reporting membership probabilities to given subgroups and distances from normality or disease; those criteria could then be used to plan treatment or quantify changes after therapy.

The ML algorithm gave "healthy" control subjects a mean probability of 0.44 for membership of cluster 2 ("diseased") (see table 2). This could be interpreted as failure of the method to adequately identify healthy subjects, but in our opinion a more likely explanation is that our asymptomatic control population, who had a median age of 67 years, already had some subclinical abnormalities; for example, although not statistically significant, the median NTpro-BNP value was slightly higher in cluster 1. None of the healthy subjects was identified by ML to have severe disease; they were mostly classified in the transition zone or as very mild HFPEF subjects. This interpretation would also imply that current diagnostic consensus criteria have limitations. To resolve such questions, much larger longitudinal studies with outcome data will be required.

Strengths and limitations

Our learning algorithm, from feature extraction to interpretation of results, was guided by pathophysiologic considerations. Analysis was focused on the LV basal regions as they capture the global longitudinal changes usually present in HFPEF subjects[28]. Secondly, we exploited all

the explanatory power of multiple high-dimensional descriptors using a previously validated unsupervised algorithm[18]. Thirdly, the multicentric data and the standardized stress protocol[11] increase the generalizability of our results.

We performed robust statistical tests to analyze our data, giving concordant results, but apart from assessing the influence of age (Data Supplement) we did not study the effect of possible confounders (gender or weight). Analysis of regional patterns, or of myocardial strain rate (relatively load-independent[16]), could also be informative. In our initial cohort of 105 healthy and HFPEF subjects, ML appeared to outperform the clinical labels. Although two observers endorsed our results by blinded reinterpretation of the echocardiographic studies (Data Supplement), no external reference is available to validate this. Invasive hemodynamic testing would have provided objective measurements of filling pressures. Our findings should be considered with caution. Larger numbers of subjects will be needed to derive more robust conclusions that could be translated into diagnostic criteria for regular clinical use.

We studied a few patients with atrial fibrillation, since it was not an exclusion criterion for the study, but with larger numbers we could independently analyze subjects in sinus rhythm and those in atrial fibrillation.

Clinical perspective

Assessing cardiac function during exercise helps to characterize the HFPEF syndrome, suggesting that diagnostic recommendations should include routine measurements of functional reserve. Diagnosis of the HFPEF syndrome needs to be refined; machine learning could help to identify subgroups with distinct phenotypes that might benefit from specific treatments, and it may offer a more reliable alternative than current diagnostic criteria.

CIRCCVIM/2017/007138 R2

Funding:

This study was supported by the European Union Seventh Framework Program (MEDIA FP7-HEALTH-2010-261409; VP2HF FP7-2013-611823), the European Research Council (MedYMA ERC-AdG-2011-291080), the Spanish Ministry of Economy and Competitiveness (TIN2014-52923-R, and the "María de Maeztu" Programme for R&D (MDM-2015-0502), Madrid, Spain), the "Fundació La Marató de TV3" (nº20154031, Barcelona, Spain), the "Programme Avenir Lyon Saint-Etienne" (IMPULSION-PALSE/2016/77, Lyon, France), and FEDER. S. Sanchez-Martinez is supported by a fellowship from "la Caixa" Banking Foundation (Barcelona, Spain).

Conflict of Interest: none declared.

References

1. Ponikowski P, Voors AA, Anker SD, Bueno H, González-Juanatey JR, Harjola V-P, Jankowska E, Jessup M, Linde C, Nihoyannopolous P, Parissis J, Pieske B, Riley J, Rosano G, Ruilope L, Ruschitzka F, Rutten F, Meer P van der. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2016;37:2129–2200.
2. Lekavich CL, Barksdale DJ, Neelon V, Wu J-R. Heart failure preserved ejection fraction (HFpEF): an integrated and strategic review. *Heart Fail Rev*. 2015;20:643–653.
3. Abbate A, Arena R, Abouzaki N, Van Tassell BW, Canada J, Shah K, Biondi-Zoccai G, Voelkel NF. Heart failure with preserved ejection fraction: refocusing on diastole. *Int J Cardiol*. 2015;179:430–440.
4. Borlaug BA, Paulus WJ. Heart failure with preserved ejection fraction: pathophysiology, diagnosis, and treatment. *Eur Heart J*. 2011;32:670–9.
5. Senni M, Paulus WJ, Gavazzi A, Fraser AG, Díez J, Solomon SD, Smiseth OA, Guazzi M, Lam CSP, Maggioni AP, Tschöpe C, Metra M, Hummel SL, Edelmann F, Ambrosio G, Stewart Coats AJ, Filippatos GS, Gheorghiade M, Anker SD, Levy D, Pfeffer MA, Stough WG, Pieske BM. New strategies for heart failure with preserved ejection fraction: the importance of targeted therapies for heart failure phenotypes. *Eur Heart J*. 2014;35:2797–2815.
6. Kao DP, Lewsey JD, Anand IS, Massie BM, Zile MR, Carson PE, McKelvie RS, Komajda M, McMurray JJ, Lindenfeld J. Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response. *Eur J Hear Fail*. 2015;17:925–935.
7. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO,

Huang C-C, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*. 2014;131:269–279.

8. Wolz R, Aljabar P, Hajnal J V, Lötjönen J, Rueckert D. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. *Med Image Anal*. 2012;16:819–30.
9. Ye DH, Desjardins B, Hamm J, Litt H, Pohl KM. Regional manifold learning for disease classification. 2014;33:1236–1247.
10. Borlaug BA, Nishimura RA, Sorajja P, Lam CSP, Redfield MM. Exercise hemodynamics enhance diagnosis of early heart failure with preserved ejection fraction. *Circ Hear Fail*. 2010;3:588–595.
11. Erdei T, Smiseth OA, Marino P, Fraser AG. A systematic review of diastolic stress tests in heart failure with preserved ejection fraction, with proposals from the EU-FP7 MEDIA study group. *Eur J Heart Fail*. 2014;16:1345–61.
12. Nedeljkovic I, Banovic M, Stepanovic J, Giga V, Djordjevic-Dikic A, Trifunovic D, Nedeljkovic M, Petrovic M, Dobric M, Dikic N, Zlatar M, Beleslin B. The combined exercise stress echocardiography and cardiopulmonary exercise test for identification of masked heart failure with preserved ejection fraction in patients with hypertension. *Eur J Prev Cardiol*. 2016;23:71–77.
13. Kosmala W, Rojek A, Przewlocka-Kosmala M, Mysiak A, Karolko B, Marwick TH. Contributions of nondiastolic factors to exercise intolerance in heart failure with preserved ejection fraction. *J Am Coll Cardiol*. 2016;67:659–670.
14. Wang J, Fang F, Wai-Kwok Yip G, Sanderson JE, Feng W, Xie JM, Luo XX, Lee APW, Lam YY. Left ventricular long-axis performance during exercise is an important prognosticator in patients with heart failure and preserved ejection fraction. *Int J Cardiol*. 2015;178:131–135.

15. Mädler CF, Payne N, Wilkenshoff U, Cohen A, Derumeaux GA, Piérard LA, Engvall J, Brodin LÅ, Sutherland GR, Fraser AG. Non-invasive diagnosis of coronary artery disease by quantitative stress echocardiography: optimal diagnostic models using off-line tissue Doppler in the MYDISE study. *Eur Heart J*. 2003;24:1584–1594.

16. Bijnens BH, Cikes M, Claus P, Sutherland GR. Velocity and deformation imaging for the assessment of myocardial dysfunction. *Eur J Echocardiogr*. 2009;10:216–26.

17. Paulus WJ, Tschöpe C, Sanderson JE, Rusconi C, Flachskampf FA, Rademakers FE, Marino P, Smiseth OA, De Keulenaer G, Leite-Moreira AF, Borbély A, Edes I, Handoko ML, Heymans S, Pezzali N, Pieske B, Dickstein K, Fraser AG, Brutsaert DL. How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. *Eur Heart J*. 2007;28:2539–50.

18. Sanchez-Martinez S, Duchateau N, Erdei T, Fraser AG, Bijnens BH, Piella G. Characterization of myocardial motion patterns by unsupervised multiple kernel learning. *Med Image Anal*. 2017;35:70–82.

19. Duchateau N, De Craene M, Piella G, Silva E, Doltra A, Sitges M, Bijnens BH, Frangi AF. A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. *Med Image Anal*. 2011;15:316–28.

20. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput*. 2003;15:1373–1396.

21. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning (Vol.1). Springer, Berlin: Springer series in statistics; 2001.

22. Kraigher-Krainer E, Shah AM, Gupta DK, Santos A, Claggett B, Pieske B, Zile MR, Voors AA, Lefkowitz MP, Packer M, McMurray JJV, Solomon SD. Impaired systolic

function by strain imaging in heart failure with preserved ejection fraction. *J Am Coll Cardiol*. 2014;63:447–456.

23. Davidavicius G, Kowalski M, Williams RI, D'Hooge J, Di Salvo G, Pierre-Justin G, Claus P, Rademakers F, Herregods MC, Fraser AG, Pierard LA, Bijnens B, Sutherland GR. Can regional strain and strain rate measurement be performed during both dobutamine and exercise echocardiography, and do regional deformation responses differ with different forms of stress testing? *J Am Soc Echocardiogr*. 2003;16:299–308.

24. Sharifov OF, Schiros CG, Aban I, Denney TS, Gupta H. Diagnostic accuracy of tissue Doppler index E/e' for evaluating left ventricular filling pressure and diastolic dysfunction/heart failure with preserved ejection fraction: A systematic review and meta-analysis. *J Am Heart Assoc*. 2016;5:e002530.

25. Suk H Il, Lee SW, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage*. 2014;101:569–582.

26. Sanchis L, Vannini L, Gabrielli L, Duchateau N, Falces C, Andrea R, Bijnens B, Sitges M. Interatrial dyssynchrony may contribute to heart failure symptoms in patients with preserved ejection fraction. *Echocardiography*. 2015;32:1655–1661.

27. Mordi IR, Singh S, Rudd A, Srinivasan J, Frenneaux M, Tzemos N, Dawson DK. Comprehensive Echocardiographic and Cardiac Magnetic Resonance Evaluation Differentiates Among Heart Failure With Preserved Ejection Fraction Patients, Hypertensive Patients, and Healthy Control Subjects. *JACC Cardiovasc Imaging*. 2017 (in press);https://doi.org/10.1016/j.jcmg.2017.05.022

28. Teske AJ, De Boeck BWL, Melman PG, Sieswerda GT, Doevendans PA, Cramer MJM. Echocardiographic quantification of myocardial function using tissue deformation imaging, a guide to image acquisition and analysis using tissue Doppler

and speckle tracking. *Cardiovasc Ultrasound*. 2007;5:27.

Legends to figures

Figure 1 (central illustration): Overview of the methods

Learning: (step#1) for each feature, definition of the pair-wise similarity between subjects; (step#2) dimensionality reduction through unsupervised learning; (step#3) output representation; (step#4) unsupervised clustering. Interpretation: (step#5) comparison of clinical indexes between clusters; (step#6) reconstruction of the variability associated with each cluster; (step#7) computation of distances and region discovery.

Extension: (step#8) New cases analysis.

Figure 2: Inputs for the machine learning

A: Velocity traces were divided into five cardiac phases, and temporally aligned to a reference. The convergent arrows indicate downsampling, and the divergent arrows indicate upsampling, to match the reference number of data points. The temporal normalization was captured by two descriptors, corresponding to the normalization of traces at rest and exercise.

B: Aligned septal and lateral velocity traces, at rest and submaximal exercise, during isovolumic contraction (IVC), systole, isovolumic relaxation (IVR), early filling, and late (atrial) filling (20 descriptors).

AC = atrial contraction; AVC = aortic valve closure; AVO = aortic valve opening; MVC = mitral valve closure; MVO = mitral valve opening.

Figure 3: Variability of learned characteristics of the clusters

Hypothetical velocity curves corresponding to the 1st and 2nd modes of the clusters identified by ML, at rest and during exercise (submax) in the basal septum and the basal lateral wall of the LV. Five curves are illustrated in each panel, representing -2 and -1 standard deviations (solid lines), the mean trace, and +1 and +2 standard

deviations (dotted lines) along each mode. The bars below each plot indicate the temporal variability in the occurrence of mitral valve closure (MVC), aortic valve opening (AVO), aortic valve closure (AVC), mitral valve opening (MVO), and onset of atrial contraction (AC); for each, the two vertical lines and the shaded area in the same color display the range from -2 to +2 standard deviations as a percentage during the cardiac cycle.

Figure 4: Variations in discriminant power between features and by clusters

The variability of 7 features identified by ML are displayed, normalized by the magnitude of each feature, with dark blue representing minimum and bright red maximum values. For each feature, 80 items are reported (2 clusters × 4 traces × first 10 modes).

Overall amplitude of the velocity profile calculated as the average of the integral of the 5 reconstructed traces per mode (mean, ±1 and ±2 standard deviations).

Diastolic fusion calculated as the average of the sum of the difference of each diastolic (early and late) negative peak to the diastasis plateau value (between the peaks).

Systolic and diastolic delays calculated as the standard deviation of the timing of systolic and early diastolic peak velocities among the 5 traces per mode.

Systolic and diastolic durations calculated as the time difference between the shortest and longest systolic and early diastolic durations calculated for the 5 traces per mode.

Atrial delay calculated as the standard deviation of the timing of the late diastolic peak (from atrial contraction) calculated for the 5 traces per mode.

Figure 5: Comparison of learned and clinically assigned diagnostic labels

Clusters distribution in the first two dimensions of the low-dimensional space identified by ML. Discordant cases are highlighted in green.

Figure 6: Distances from each subject to the center of each cluster

(A) All healthy and HFPEF subjects (according to clinical labels) displayed by their distances from clusters 1 and 2 (*healthy* and *diseased* clusters identified by ML). Cases with discordant clinical and ML labels are highlighted in green, and the probabilities of membership to each cluster are indicated by dashed lines; the blue, red and green areas correspond to healthy, transition, and HFPEF zones. (B) and (C) display hypertensive and breathless controls, mapped using the algorithm learned from the analysis of the groups shown in (A).

Table 1: **Comparisons between groups**

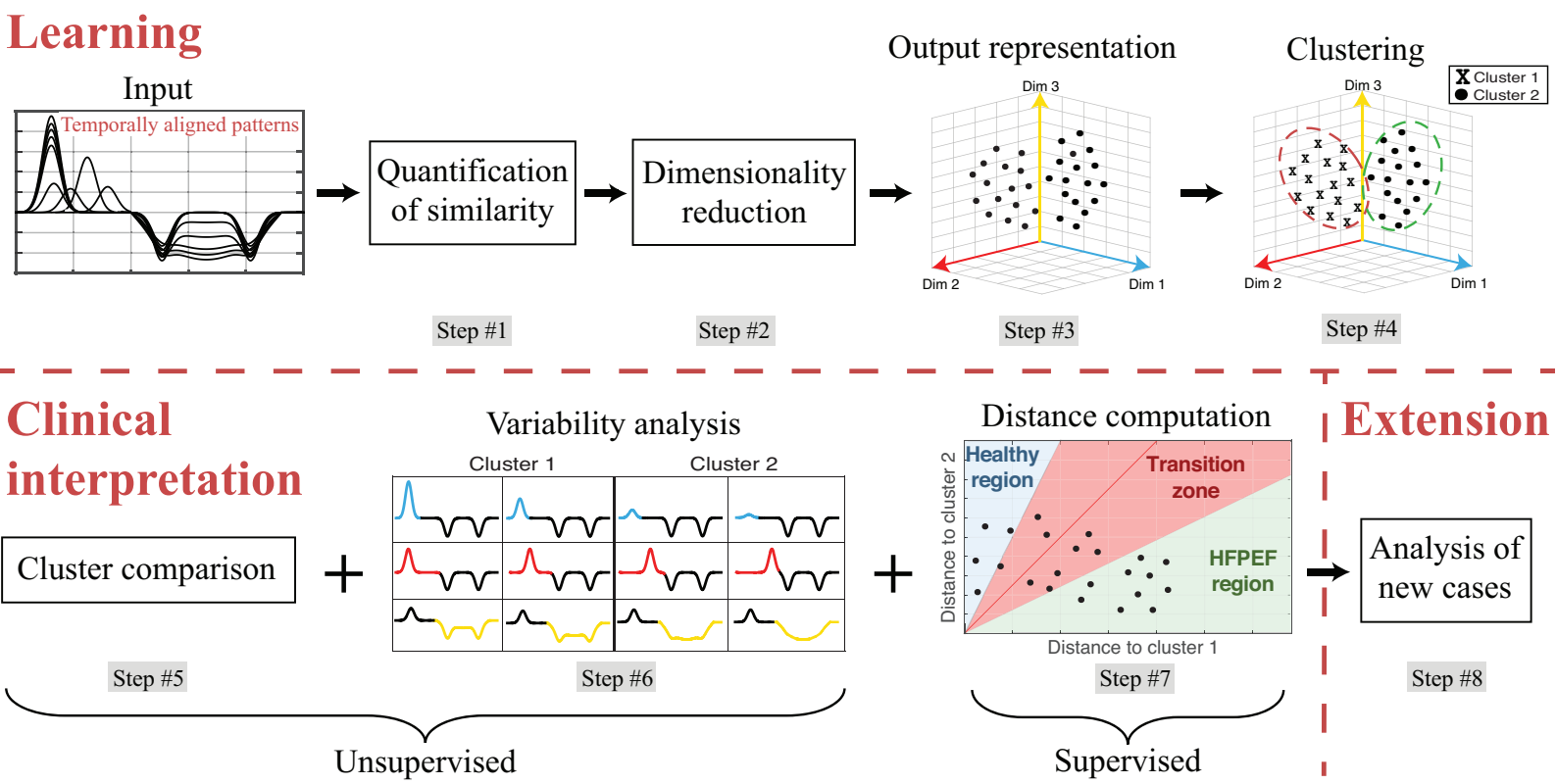
	Healthy (n=33)	Cluster 1 (n=26)	p-value	HFPEF (n=72)	Cluster 2 (n=79)	p-value	Cluster 1 vs. 2 p-value
Age (y)	66.9(64–69.1)	67.02(63–70.6)	0.81	72(68.0–78.0)	71(67–77)	0.39	**0.005**
Female, n (%)	20(60.6)	18(69.2)	0.68	51(70.8)	53(67.1)	0.78	0.84
Caucasian race, n (%)	32(97.0)	26(100)	0.06	71(98.6)	77(97.5)	0.30	0.09
Body mass index (kg/m^2)	25.3(23.2–28.8)	24.8(23.2–29.0)	0.99	28.8(25.8–32.8)	28.1(25.4–31.6)	0.58	**0.004**
6-minute walk test (m)	501(476–560)	555(465–565)	0.64	357(305–395)	385(330–470)	0.09	**0.001**
N-terminal B-type natriuretic peptide (ng/mL)	70(31–119)	75(48–154)	0.20	220(87–330)	139(64–325)	0.20	**0.049**
E/e' ratio – Rest	6.9(5.9–8.6)	8.5(6.7–11.1)	0.13	10.8(8.6–13.7)	9.3(7.8–13.3)	0.13	**0.03**
E/e' ratio – Submax	8.1(6.1–9.3)	9.2(7.8–10.2)	**0.048**	10.8(8.7–13.8)	10.2(7.7–11.8)	0.07	0.45
E/A ratio – Rest	1.00(0.84–1.21)	0.93(0.79–1.20)	0.58	0.88(0.77–1.05)	0.90(0.79–1.07)	0.47	0.45
E/A ratio – Submax	1.04(0.90–1.23)	1.05(0.90–1.26)	0.64	1.06(0.86–1.20)	1.04(0.87–1.19)	0.81	0.43
LV ejection fraction (%)	62.6(60.4–64.7)	62.1(60.6–64.2)	0.78	60.6(57.0–63.9)	60.8(57.1–64.8)	0.77	0.09
LV mass index (g/m^2)	72.7(60.8–84.9)	81.5(64.0–90.8)	0.24	108.5(93.0–132.2)	104.6(88.3–127.7)	0.34	**0.00002**
Deceleration time – rest (ms)	230(201–261)	237(219–265)	0.43	236(188–272)	233(192–272)	0.75	0.39
Deceleration time – submax (ms)	152(135–166)	153(135–180)	0.95	156(136–190)	157(137–182)	0.78	0.69
LV end-diastolic volume index (mL/ m^2)	44.6(37.1–54.0)	52.6(38.8–59.8)	0.27	46.9(38.0–59.5)	44.9(37.3–56.3)	0.54	0.25
Ard-Ad (ms)	-7(-20–2)	–8(-21–2)	0.99	-9(-20–6)	–10(-20–6)	0.97	0.83
Left atrial volume index (mL/m^2)	24.7(21.0–34.4)	34.1(23.2–39.0)	0.06	37.4(33.5–44.6)	35.7(27.6–42.6)	0.12	0.15

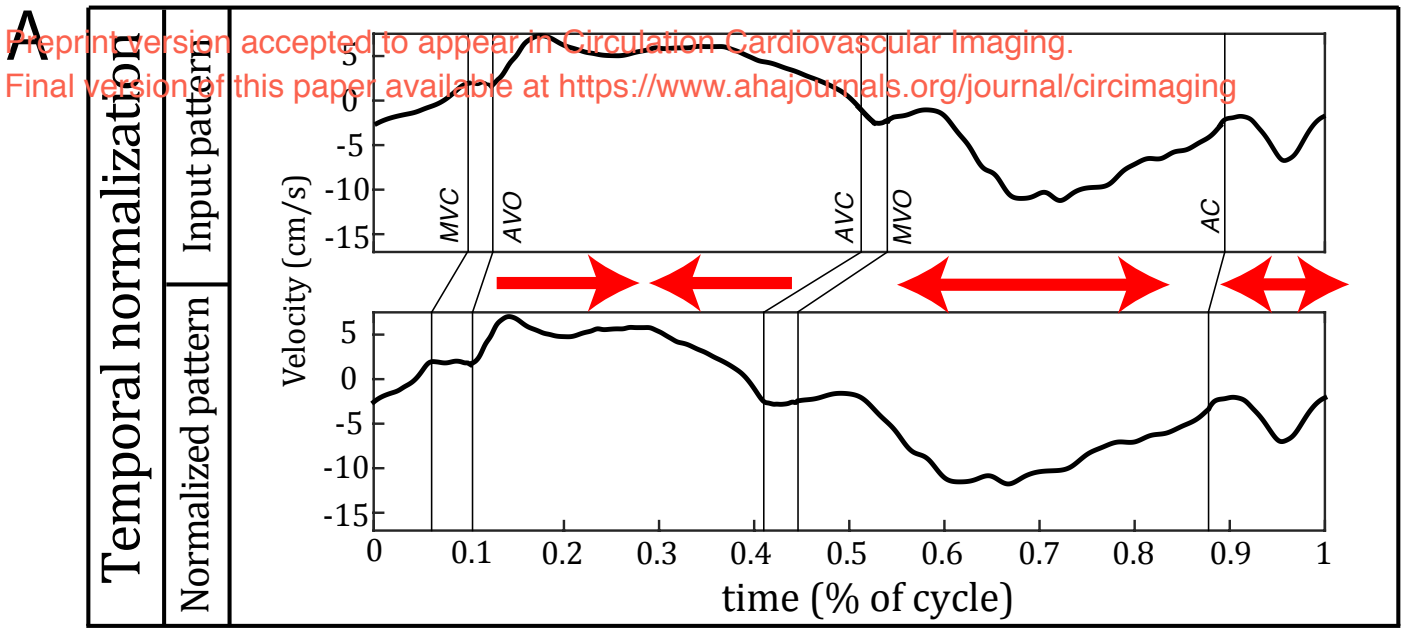
Categorical variables expressed as counts and percentages. Continuous variables expressed as median (25th–75th percentile). A-a = duration of mitral valve flow during atrial contraction minus duration of pulmonary vein retrograde flow; E/A = ratio of the early and late transmitral flow velocities; E/e' = ratio of the early transmitral flow velocity and the early diastolic mitral annular velocity; LV = left ventricular; NS = non-significant.

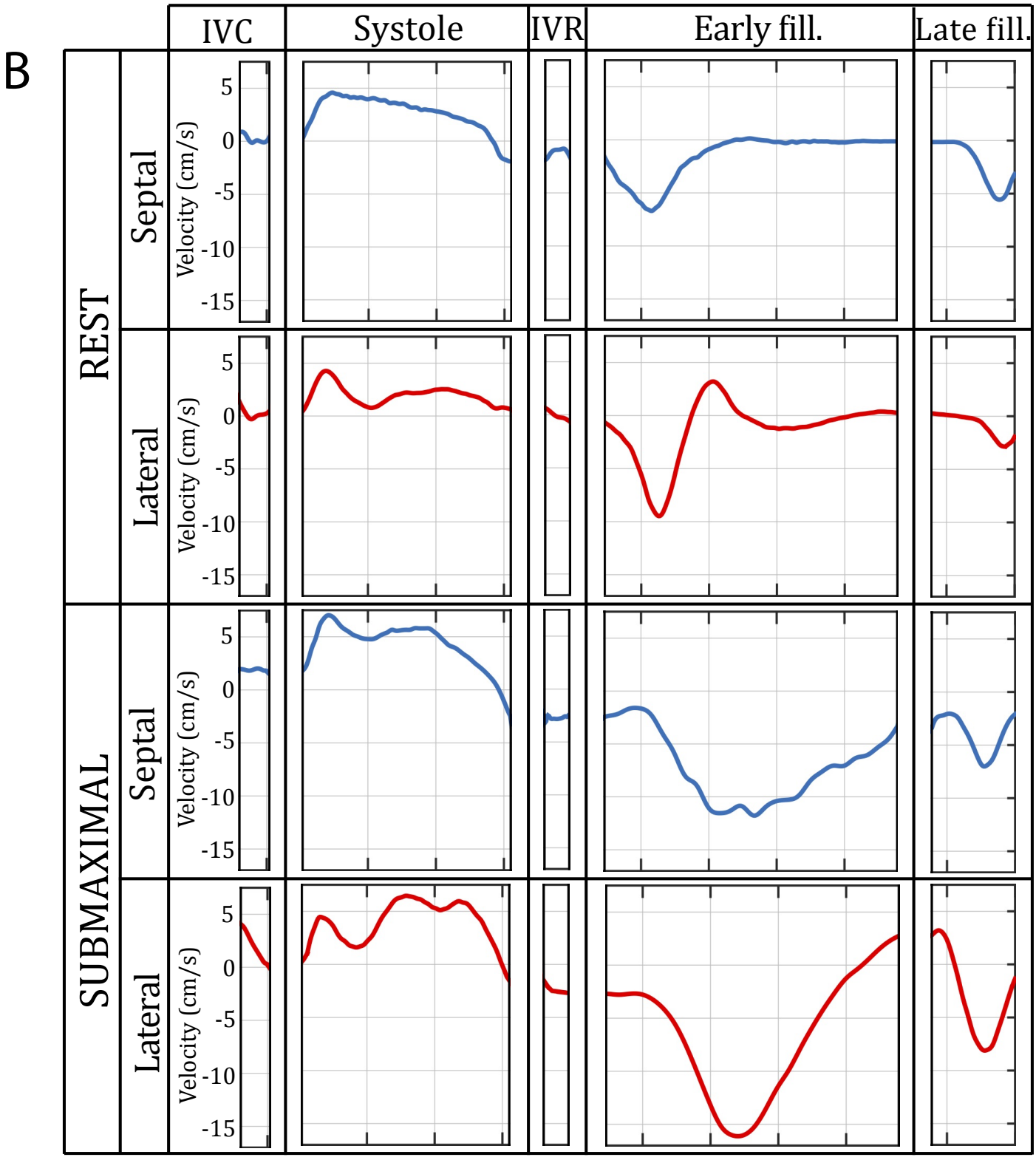
Table 2: **Comparison of clinical and learned classifications of subjects**

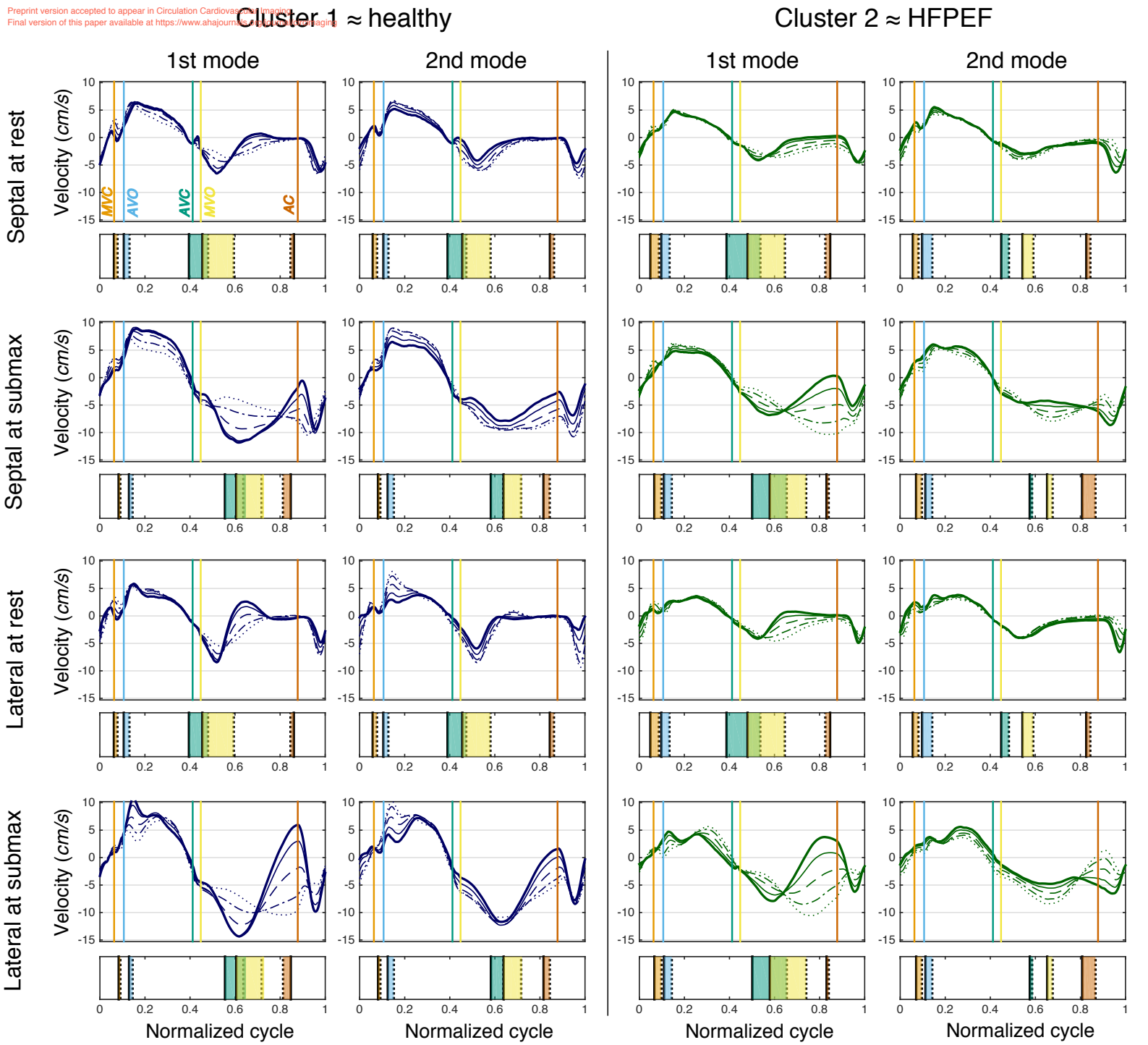
		Classification by machine learning					
		Definite healthy (n=15)	**Transition zone: possibly normal (n=21)**	**Transition zone: possibly HFPEF (n=41)**	**Definite HFPEF (n=79)**	**Membership of cluster 1**	**Membership of cluster 2**
Clinical labels	**Healthy (n=33)**	13(39.4%) True Negatives	5(15.1%)	13(39.4%)	2(6.1%)	0.56	0.44
	Hypertensive (n=24)	0(0%)	6(25.0%)	10(41.7%)	8(33.3%)	0.38	0.62
	Breathless (n=27)	1(3.7%)	4(14.8%)	4(14.8%)	18(66.7%)	0.30	0.70
	HFPEF (n=72)	1(1.4%)	6(8.3%)	14(19.4%)	51(70.8%) True Positives	0.25	0.75

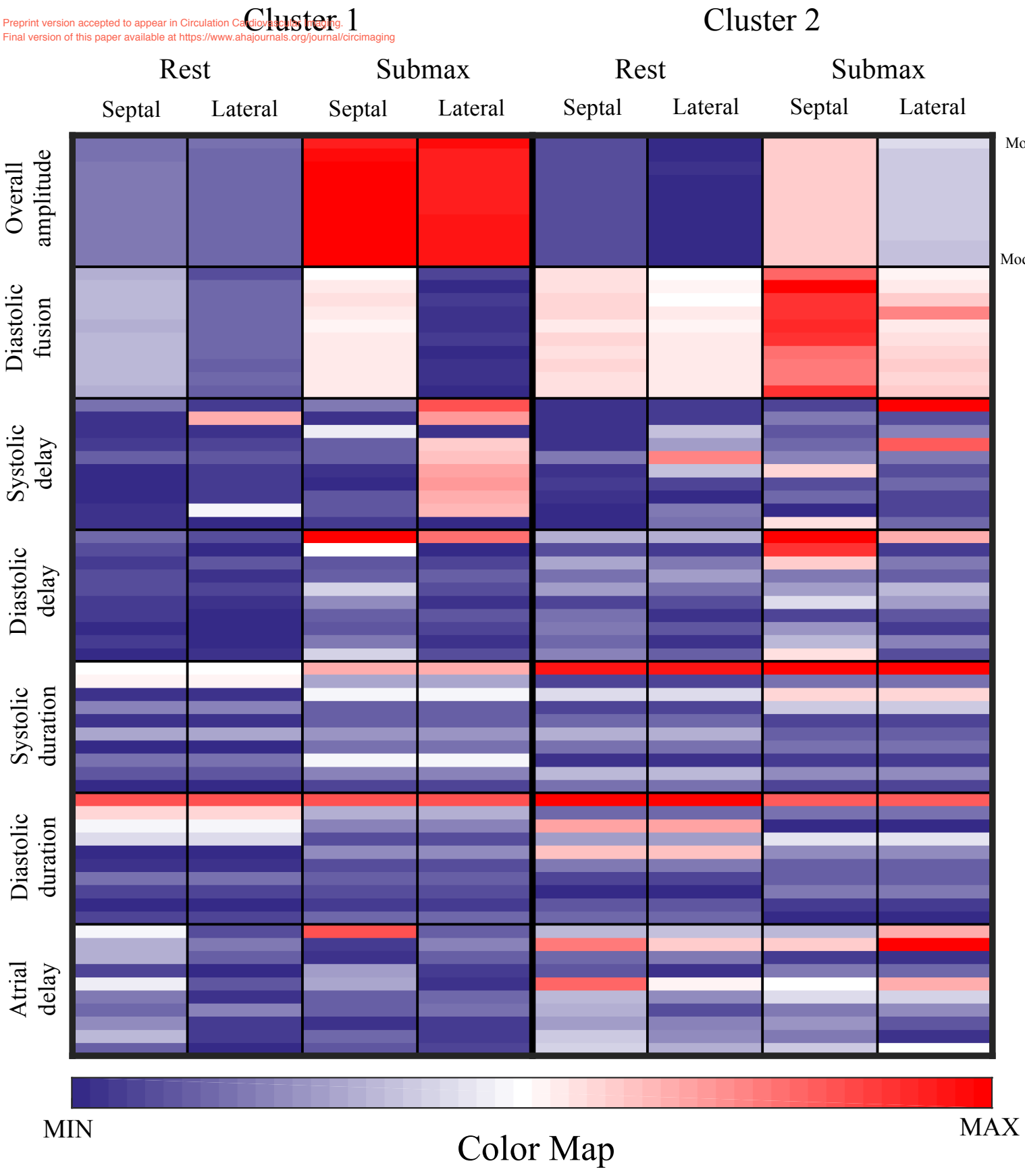
The left quadrant contains the confusion matrix for the clinical labels compared to the classification by ML. The right quadrant summarizes the mean probabilities of each of the clinical groups of belonging to the clusters identified by ML.

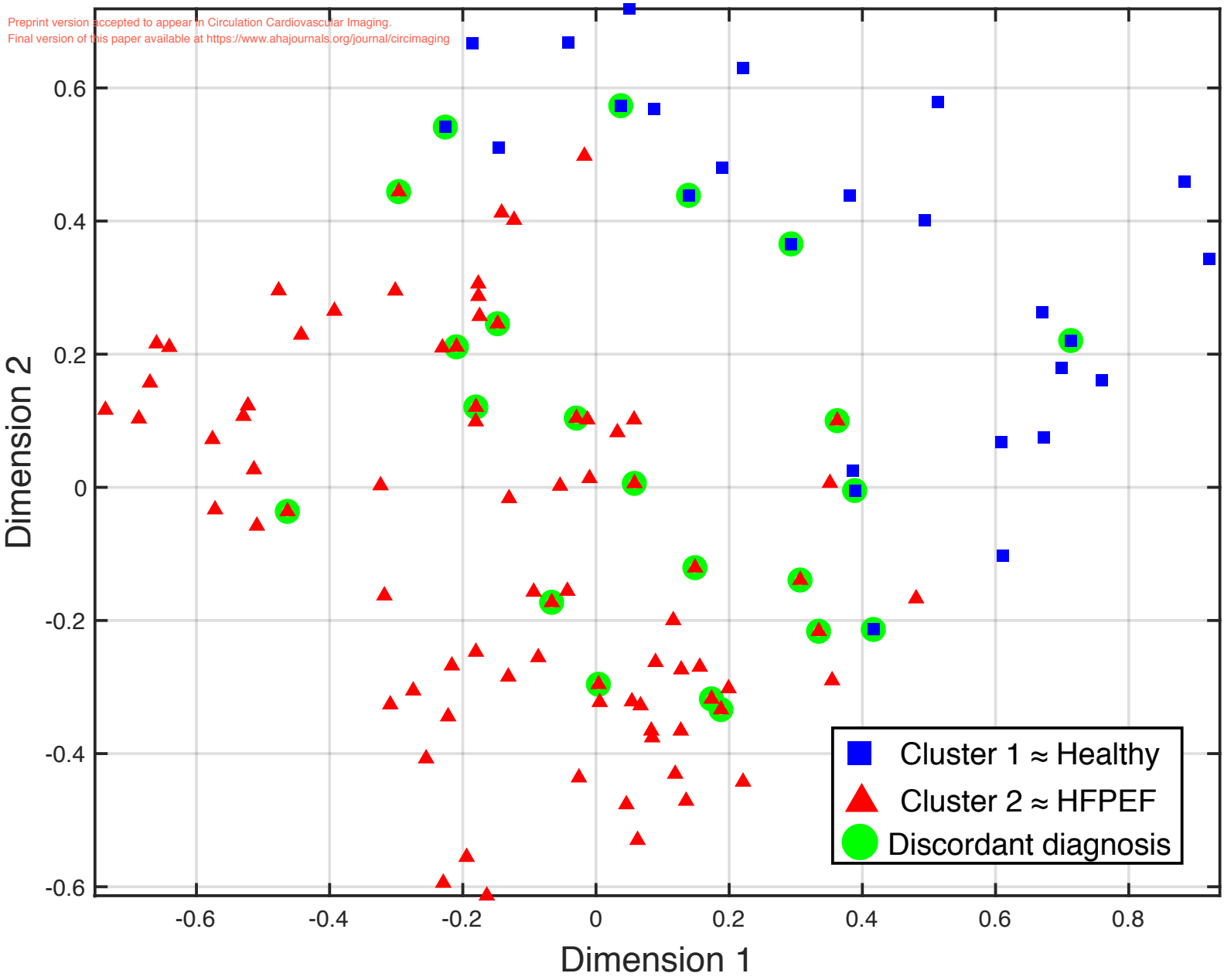


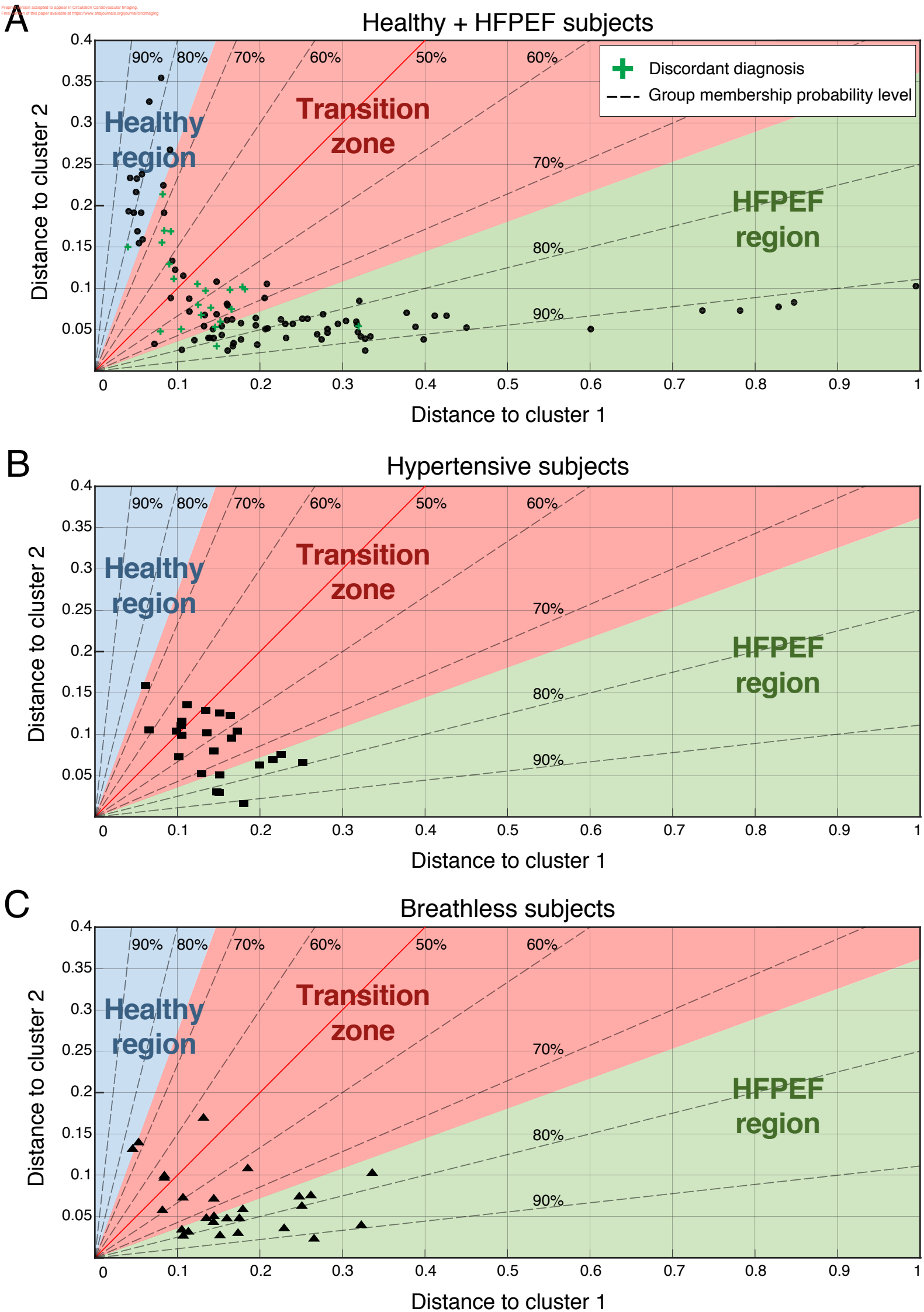












SUPPLEMENTAL MATERIAL

Re-analysis of diagnostic images

Methods:

All discordant cases out of the 105 subjects analyzed were reinvestigated by detailed study of their stored echocardiographic images. Two experienced observers reviewed 36 studies independently, including also 3 healthy ("true negatives") and 3 HFPEF subjects with concordant diagnoses ("true positives"), and 8 breathless subjects lying within the transition zone or the HFPEF region. The observers were blinded for this exercise since all studies were presented anonymized and in random order.

Results:

Blinded re-interpretation of the echocardiographic data of discordant diagnosis cases revealed possible explanations in most cases. Nine out of fifteen "healthy" subjects who did not map to the healthy zone on ML, had focal hypertrophy of the outlet ventricular septum or borderline LV hypertrophy. Other findings in this group included possible apical hypertrophic cardiomyopathy (1 subject), silent myocardial ischemia during exercise (1 subject), and right ventricular diastolic dysfunction (1 subject).

Findings in subjects diagnosed as HFPEF but allocated to other zones by ML included hypertensive heart disease with preserved functional reserve (3 subjects), and left bundle branch block with dyssynchrony (1 subject).

On review, the breathless subjects located in the transition zone had either no definite abnormality or mild hypertensive heart disease, while those in the HFPEF region showed hypertensive heart disease, inducible ischemia, or impaired right ventricular function.

Discussion:

Re-analysis of the discordant cases with segmental interrogation of strain and strain rate

was particularly informative. For example, some subjects who were identified as abnormal by ML, had been recruited as healthy controls − implying normal blood pressure. Nonetheless they had septal hypertrophy; recognized as an early sign of hypertensive heart disease [1]. These individuals might have occult hypertension [2], or increased late systolic loading from wave reflections related to central arterial stiffness [3]. In either case, while their LV long-axis motion might be affected, they would not be diagnosed using routine clinical tests but they could be unmasked by ML that detected impaired LV long-axis functional reserve.

References

1. Baltabaeva A, Marciniak M, Bijnens B, Moggridge J, He FJ, Antonios TF, MacGregor GA, Sutherland GR. Regional left ventricular deformation and geometry analysis provides insights in myocardial remodelling in mild to moderate hypertension. *Eur J Echocardiogr*. 2008;9:501–508.
2. Gaudron PD, Liu D, Scholz F, Hu K, Florescu C, Herrmann S, Bijnens B, Ertl G, Störk S, Weidemann F. The septal bulge - an early echocardiographic sign in hypertensive heart disease. *J Am Soc Hypertens*. 2016;10:70–80.
3. Chirinos JA, Kips JG, Jacobs DR, Brumback L, Duprez DA, Kronmal R, Bluemke DA, Townsend RR, Vermeersch S, Segers P. Arterial wave reflections and incident cardiovascular events and heart failure: MESA (Multiethnic Study of Atherosclerosis). *J Am Coll Cardiol*. 2012;60:2170–2177.

Influence of age

Since subjects who were recruited as HFPEF were slightly older than the healthy volunteers, the unsupervised ML algorithm was repeated with age as an additional input. This had a negligible impact on the final characterization; the correlation between the two distributions was 0.97. Only 3 of the 22 subjects with discordant clinical and ML diagnoses were reclassified, and they were close to the frontier between clusters, which highlights their uncertain diagnosis.

Supplementary table S1: Comparison among clinical centers

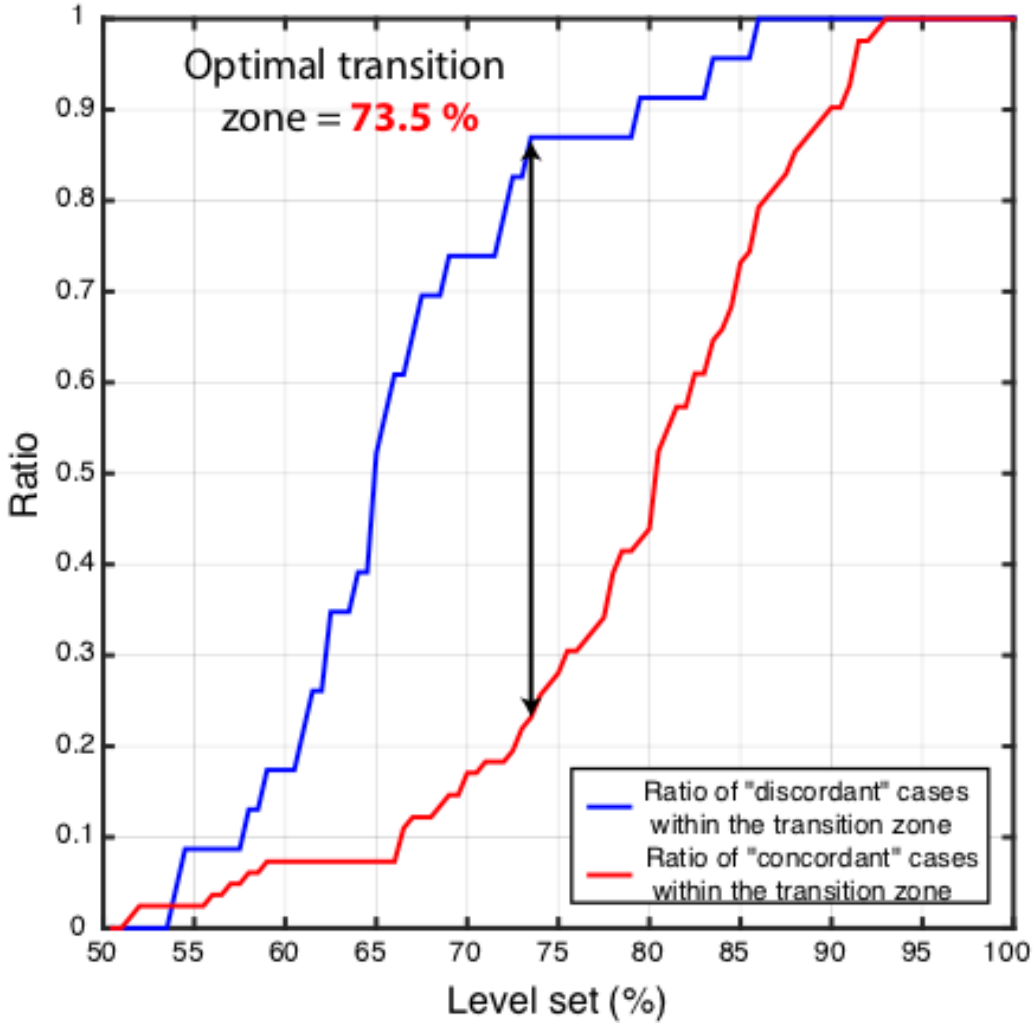
The table shows the distribution of subjects between the four groups, across the four participating centers. Mean ages for all groups were about 70 years, and all subjects apart from 2 were Caucasian. The female/male ratio was similar between centers, except that subjects studied in Oslo were predominantly female. On average, all groups were overweight.

	Cardiff (n=76)				Novara (n=31)				Perugia (n=20)				Oslo (n=29)			
Class	H	HT	B	HF	H	HT	B	HF	H	HT	B	HF	H	HT	B	HF
Number of subjects	23	15	15	23	-	4	12	15	-	5	-	15	10	-	-	19
Age, y	69.6 ± 6.6				66.1 ± 7.4				71.4 ± 5.6				71.4 ± 5.5			
Female n (%)	43 (56.6)				18 (58.1)				12 (60)				24 (82.8)			
Caucasian n (%)	74 (97.4)				31(100)				20 (100)				29 (100)			
BMI (kg/m^2)	29.1 ± 5.4				27.3 ± 3.7				28.4 ± 3.6				26.4 ± 4.6			

Categorical variables are expressed as counts and percentages and continuous variables are expressed as mean ± SD. H, HT, B and HF = healthy, hypertensive, breathless and HFPEF, respectively.

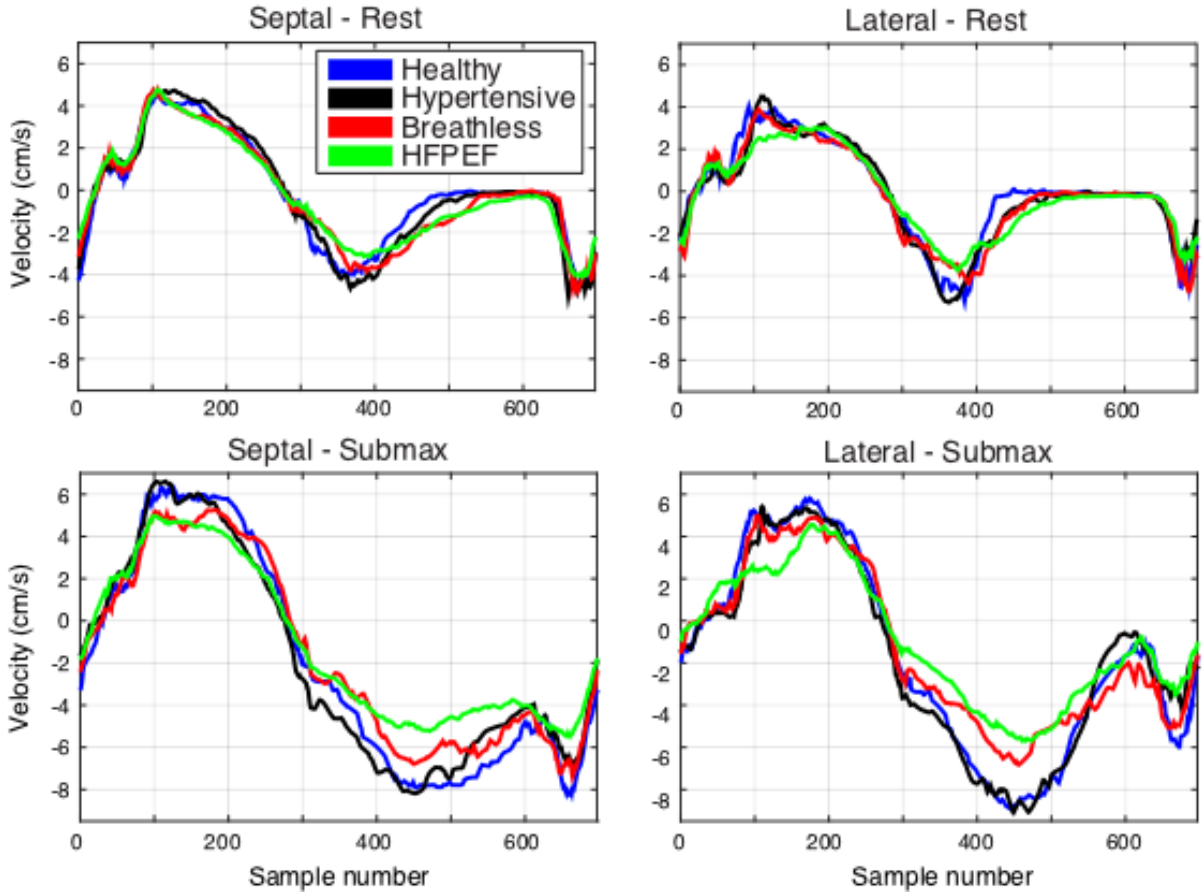
<u>Supplementary figure S1</u>: Definition of the cut-points between diagnostic zones

To define the cut-points between the healthy/HFPEF regions and the intermediate transition zone, we gradually incremented the area of the transition zone with steps of 0.5%, ranging from 50% (red line in Fig.6) to 100% (X axis = extreme healthy; Y axis = extreme HFPEF). We calculated the ratio of "discordant" and "concordant" cases within the transition zone for each of the tested configurations, and chose the one that maximized the ratio difference.



<u>Supplementary figure S2</u>: Comparisons between groups

The traces correspond to the median basal septal and lateral myocardial velocity profiles measured throughout the cardiac cycle during rest and submaximal exercise. Results are shown for each of the groups analyzed.



Supplementary video: Clusters 1 and 2 separated by the first 3 dimensions

3D rotating figure of the clusters plotted in the first 3 output space dimensions. Subjects belonging to cluster 1 (≈ healthy) are indicated by a blue square. Subjects belonging to cluster 2 (≈ HFPEF) are indicated by a red triangle. Discordant diagnoses cases are highlighted with a green circle.