This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/149832/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:

Fraser, Alan G., Tschöpe, Carsten and Boer, Rudolf A. 2022. Diagnostic recommendations and phenotyping for heart failure with preserved ejection fraction: knowing more and understanding less? European Journal of Heart Failure 23 (6), pp. 964-972. 10.1002/ejhf.2205 file

Publishers page: http://dx.doi.org/10.1002/ejhf.2205
<http://dx.doi.org/10.1002/ejhf.2205>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Diagnostic recommendations and phenotyping for heart failure with preserved ejection fraction: knowing more and understanding less?

Alan G. Fraser1,2,3*, Carsten Tschöpe4,5, and Rudolf A. de Boer6

1 School of Medicine, Cardiff University, Cardiff, UK; 2 Department of Cardiology, University Hospital of Wales, Cardiff, UK; 3 Cardiovascular Imaging and Dynamics, Catholic University of Leuven, Leuven, Belgium; 4 Berlin Institute of Health at Charité (BIH), Universitätmedizin Berlin, BIH Center for Regenerative Therapies (BCRT), Berlin, Germany; 5 Department of Cardiology, Campus Virchow (CVK), Charité Universitätmedizin Berlin, Berlin, Germany; and 6 Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

In 1982 John Goodwin was discussing hypertrophic cardiomyopathy1 but his statement could apply even more to the ever-expanding tangle of diagnostic and prognostic criteria, classifications, scores, and phenotypes in which we are now enmeshed when considering heart failure (HF) in patients who have a normal left ventricular ejection fraction (EF). The syndrome was known as diastolic HF2,3 before it became HF with normal EF.4 In 2005 the guideline task force of the European Society of Cardiology (ESC) referred to preserved left ventricular EF (PLVEF)3 after the term had been introduced by the CHARM-Preserved clinical trialists in 2003.5 ‘Heart failure with preserved ejection fraction’ (HFpEF) was then adopted by the ESC guidelines in 2008,6 since when that label has been retained7,8 (Table 1). The cut-point for a normal EF has been kept at >50% without any consensus statement having cited a normative database, and the same shortcoming has been maintained even in the most recent international consensus that proposes four stages and four types of HF.9 Instead, the choice has been based on ‘historical’7,8 or ‘traditional’9 grounds. Unvalidated variations have also been made in consecutive recommendations for assessing diastolic function, with major impact on prevalence but without evidence of improved performance or utility.10

There is growing appreciation that this is not a trivial issue: imprecision of diagnostic criteria for HFpEF has contributed to the heterogeneity of patients recruited into therapeutic trials and to the preponderance of negative outcomes. Rather than reassessing the recommendations completely, however, additional criteria were added such as natriuretic peptides in 20074 and left ventricular volume in 2012.7 The concept of HF with mid-range ejection fraction (HFmrEF) was introduced in 20168 in recognition that subjects around the arbitrary cut-point for EF were being excluded from trials, but EF is a continuously distributed variable that alone is neither sufficient to indicate cardiac output nor predictive of left ventricular filling pressures. Diagnostic utility of the HFA-PEFF scoreThe most recent ESC consensus statement for diagnosing HFpEF attempted to introduce more concordance by proposing a scoring system for diastolic indices, structural changes, and biomarkers.11 The HFA-PEFF score was developed after an extensive review of the literature but it had not been validated before publication. An alternative diagnostic score called H2FPEF was derived from one cohort and tested in another, using a pulmonary capillary wedge pressure (PCWP) ≥15 mmHg at rest or ≥25 mmHg during exercise as the reference; its C-statistic was 0.84.12 At least five studies of the diagnostic performance of the HFA-PEFF score have now been published.13–17 Barandiarán Aizpurua et al.13 reported that it can rule in HFpEF with very high specificity (93%) and positive predictive value (98%) but they determined its accuracy by evaluating patients selected by some of the same factors incorporated in the HFA-PEFF score.
more realistic evaluation tested the score against invasive haemodynamic studies. Compared with HfPEF defined as a PCWP≥15 mmHg either at rest or during exercise, the C-statistic of the HFA-PEFF score was 0.73.14 The mean age of subjects in that study was only 59 years, and 23% of 156 individuals were misclassified. In an older cohort, 72% of 286 subjects aged ≥75 years with a history of hospitalisation for HF and an EF≥50%, were classified by the HFA-PEFF score as having an intermediate probability.15 The authors suggested that the HFA-PEFF score may identify early disease, but another possibility is that it may over diagnose HfPEF in the elderly because the criteria in the score are not adjusted for age. The largest reported evaluation of the diagnostic performance of the HFA-PEFF score was performed in patients and matched controls free of cardiovascular disease who had unexplained dyspnoea. They had participated in the TOPCAT and RELAX trials, in which case they were presumed to have HfPEF, or in the ARIC study, in which case they were presumed not to have HF. A low HFA-PEFF score was reported to rule out HfPEF with extremely high sensitivity (99.5%) and negative predictive value (95.7%), while a high HFA-PEFF score ruled in HfPEF with good specificity (82.8%) and positive predictive value (79.9%).16 That evaluation, however, was also made in a group with high pre-test probability. Other investigators have tested the HFA-PEFF score against indices of cardiac functional reserve. Faxen et al.17 reported that the score was unrelated either to coronary flow reserve or to the 6-min walk test distance. They also found differences in the mean HFA-PEFF score between countries. Parcha et al.16 reported that the score did not correlate with peak oxygen consumption.

Prognostic utility of the HFA-PEFF and H2FPEF scores

Although the HFA-PEFF and H2FPEF algorithms were both developed for primary diagnosis, at least five studies have now compared their utility instead to estimate prognosis.16,18–21 In an elegant and very thorough retrospective cohort study, Verbrugge et al.18 included 443 consecutive patients with EF≥50% who had been hospitalised and treated with intravenous diuretics for acute HF. Patients with an identifiable specific aetiology such as ischaemia or valve disease were excluded. Their mean age was 78 years and they constituted a high-risk group since 69% died during an average follow-up of 28 months. Increasing values of each score similarly and strongly predicted increased risk of mortality. Intra-individual variations between the scores are not reported. Sotomi et al.19 reported similar performance of the HFA-PEFF score in 804 patients after 1 year while Selvaraj et al.20 investigated a population with a much lower pre-test probability of HF. In 641 participants aged 67–90 years with unexplained dyspnoea, 11% were judged high-risk by the H2FPEF score and 26% were high-risk according to the HFA-PEFF score. The overall prognostic power of each score was good, but only 27 subjects (4%) were designated to be at high risk by both scores while 28% had discordant findings.

In the MEDIA study, Huttin et al.21 evaluated both scores in 515 subjects with HfPEF according to the 2007 ESC consensus criteria (Table 1). They proposed a new MEDIA echo score incorporating four simple echocardiographic measurements, which together with clinical variables and natriuretic peptides had good predictive power for mortality and recurrent hospitalisation at 1 year, both in the derivation cohort (C-statistic 0.78) and when retested in the KaRen study (C-statistic 0.69). Adding the MEDIA score increased the C-statistics of both the HFA-PEFF and H2FPEF scores in the same group by 0.09 and 0.12 respectively. Before implementing these new algorithms widely, it is salutary to note that the MAGGIC (Meta-Analysis Global Group in Chronic Heart Failure) score, using clinical variables and including EF as the only imaging parameter, has similar prognostic value in HfPEF. In 407 patients who had required diuretics during an acute admission with a clinical diagnosis of HF confirmed by Framingham criteria, and who had a normal EF and raised brain natriuretic peptide, the MAGGIC score predicted outcomes at a mean follow-up of 3.6 years with C-statistics of 0.74 for
mortality and 0.66 for recurrent cardiovascular hospitalisation. The HFA-PEFF and H2FPEF algorithms may diagnose different types of patients yet fortuitously have similar predictive power. The H2FPEF score is weighted towards atrial fibrillation, observed in 34% of its derivation cohort, whereas the HFA-PEFF score includes measurements of left ventricular long-axis function and natriuretic peptides. Unsurprisingly, both perform less well in people with low or intermediate likelihood of HFpEF – when they would be most useful. There is little need for a score that identifies patients in whom a diagnosis of HF can be made by simpler criteria, or for a score that predicts outcomes without indicating how they might be changed. And of course, performance as a prognostic score does not prove causality.

Phenotyping by artificial intelligence

The optimal diagnostic and prognostic criteria for HF may differ and vary by aetiology, whether the HF is acute or chronic, and whether the EF is normal or reduced. Trying to lump together all patients with the syndrome of HFpEF according to a single diagnostic algorithm has not been conspicuously successful in identifying effective treatments, so it is illogical to refine diagnostic criteria without evidence that their application leads to better outcomes. An alternative approach that has become extremely popular is to use machine learning to explore specific phenotypes of HFpEF and diastolic function, in the hope of uncovering causative mechanisms of disease for which targeted treatment can be developed – but so far, the trend may be adding to our confusion rather than resolving it. We have identified at least 14 reports of machine learning used to investigate patients with HFpEF. They show considerable diversity of design and inclusion criteria and more variability of input data, as summarised in Table 2. Three studies that applied the modelling technique of latent class analysis for a similar objective are also listed. Most studies had no control groups and only two included data collected during exercise. None integrated data from all potential sources including demographic and clinical variables, biomarkers and proteomics, structural and functional imaging at rest and during exercise, and genomics, in a population at risk. A minority of studies retested their findings in an independent population. Different studies have identified between two and six phenogroups of HFpEF. Many relate to known risk factors and elements of HFpEF pathophysiology (Table 3 and Figure 1), while some give new insights. Most studies have identified phenotypes that predict varying outcomes. A few provide the first hints that this approach might identify phenogroups with differential responses to drugs. For example, TOPCAT participants with low H2FPEF scores (≤6) were more likely to benefit from spironolactone (hazard ratio 0.47), while subjects in phenogroup 3 from a study in China, who had a high prevalence of ischaemic heart disease and type 2 diabetes, had a lower mortality and fewer hospitalisations if they were taking a beta-blocker or angiotensin-converting enzyme inhibitor (absolute risk reductions >10% at 5 years).

Building a bridge to understanding

Diagnostic standards are indeed important, since using dissimilar criteria to select patients for clinical trials has a major impact. Knowledge of HFpEF is advancing but we need concerted action to improve our understanding and develop new treatments. Unfortunately, none of the consensus recommendations has really been based on a prospective evaluation and evidence of beneficial impact, whether for diagnosis, prognosis, or choice of treatment. Clinical trialists need to bypass too simple diagnostic recommendations and embrace more detailed characterisation of subjects as they are recruited for new studies. Some less common phenotypes such as amyloidosis and haemochromatosis are rather mono factorial and already amenable to specific treatments. For more complex phenotypes, we should reconsider predefined and often composite end points as the only outcomes that can be accepted. What about a really large ‘allcomers’ HF randomised controlled
trial, with fully characterised subjects selected not by EF but because of dyspnoea and reduced exercise capacity, or other independent criteria, and with an adaptive design? Then our diagnostic guidelines and our prognostic scores for HFpEF really would be useful.

References


2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2016;18:891–975.


<table>
<thead>
<tr>
<th>Study</th>
<th>Diagnostic criteria</th>
<th>n</th>
<th>M ± SD</th>
<th>Age, years</th>
<th>Machine learning method</th>
<th>Input variables</th>
<th>Photo scores</th>
<th>Prognostic score</th>
<th>Independent validation, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun 2016</td>
<td>Hospitalization, BMI, Framingham, ejection fraction, dyslipidemia</td>
<td>397</td>
<td>39 ± 12</td>
<td>65 ± 12</td>
<td>Hierarchical clustering</td>
<td>48 (drug, VHR, CHF, echocardiography)</td>
<td>4</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>Kao 2015</td>
<td>EF &lt; 40%, NYHA class 2, hypertension, BMI, Framingham, cardiac function, PRESERVE</td>
<td>4113</td>
<td>40 ± 10</td>
<td>67 ± 11</td>
<td>Linear regression</td>
<td>11 (age, risk, BMI, LV ejection fraction, CHF, CHF, echocardiography)</td>
<td>4</td>
<td>Yes</td>
<td>5</td>
</tr>
<tr>
<td>Ghan 2015</td>
<td>Heart failure</td>
<td>795</td>
<td>72 ± 10</td>
<td>64 ± 16</td>
<td>Linear regression</td>
<td>16 (age, risk, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>4</td>
<td>No</td>
<td>3</td>
</tr>
<tr>
<td>Ahmad 2019</td>
<td>CHF, New York Heart Association, Non-CHA2DS2-VASc score</td>
<td>9913</td>
<td>45 ± 10</td>
<td>67 ± 10</td>
<td>Random forest regression</td>
<td>8 (age, creatinine, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>4</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>Santamour 2018</td>
<td>ESC criteria 2017, NICE-CHD risk score</td>
<td>192</td>
<td>72 ± 11</td>
<td>72 (66-78)</td>
<td>Multiple kernel learning</td>
<td>3 (ECG, echocardiography)</td>
<td>2</td>
<td>Yes</td>
<td>51</td>
</tr>
<tr>
<td>Tasso 2018</td>
<td>ESC criteria 2017, NICE-CHD risk score</td>
<td>352</td>
<td>49 ± 7</td>
<td>64 ± 7</td>
<td>k-means clustering</td>
<td>9 (age, gender, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>2</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Hartless 2019</td>
<td>Hospitalization for heart failure</td>
<td>345</td>
<td>53 ± 14</td>
<td>72 ± 14</td>
<td>Non-parametric methods</td>
<td>2 (age, risk, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>2</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Prakash 2016</td>
<td>ESC criteria 2017, mid disease hypertension</td>
<td>177</td>
<td>77 ± 13</td>
<td>60 ± 3</td>
<td>Hierarchical clustering</td>
<td>3 (age, gender, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>2</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Cohen 2021</td>
<td>TOPCAT participants (HT, diabetes)</td>
<td>548</td>
<td>69 ± 10</td>
<td>69 ± 10</td>
<td>Linear regression</td>
<td>3 (age, BMI, LV ejection fraction, CHF, echocardiography)</td>
<td>3</td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>Filbo 2020</td>
<td>TOPCAT participants in America</td>
<td>156</td>
<td>71 ± 10</td>
<td>71 ± 10</td>
<td>Linear regression</td>
<td>11 (clinical variables)</td>
<td>3</td>
<td>No</td>
<td>2</td>
</tr>
</tbody>
</table>
Characterisation of HfPEF phenogroups

<table>
<thead>
<tr>
<th>RISK FACTORS</th>
<th>MACHINE LEARNING STUDIES</th>
<th>DISCRIMINANTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Demographic &amp; clinical variables</td>
<td>age</td>
</tr>
<tr>
<td>female sex</td>
<td></td>
<td>renal function</td>
</tr>
<tr>
<td>obesity</td>
<td></td>
<td>body mass index</td>
</tr>
<tr>
<td>hypertension</td>
<td></td>
<td>blood glucose</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td></td>
<td>hematocrit</td>
</tr>
<tr>
<td>renal dysfuncation</td>
<td></td>
<td>gender</td>
</tr>
<tr>
<td>anemia</td>
<td></td>
<td>blood pressure</td>
</tr>
<tr>
<td>iron deficiency</td>
<td></td>
<td>atrial fibrillation</td>
</tr>
<tr>
<td>sleep disorders</td>
<td></td>
<td>NYHA classification</td>
</tr>
<tr>
<td>COPD</td>
<td></td>
<td>natriuretic peptides</td>
</tr>
</tbody>
</table>

**Total N = 18 studies**

*Known risk factors for heart failure with preserved ejection fraction (HfPEF) are listed in the box on the left, in their order in the European Society of Cardiology consensus recommendations. The three diagrams in the middle box show which broad types of variables were used as input to the machine learning studies, as summarised in Table 3. The numbers refer to studies (from a maximum of 18) that studied each combination of factor. Variables that were found to discriminate between phenogroups of HfPEF are listed in the box on the right, in order of their predominance; the list includes all those that were reported by 50% or more of the studies. The commonest imaging variables (not shown) were left ventricular hypertrophy, the E/e' index, and left atrial volume, all reported by 85% of studies. Thus in general, the machine learning models have confirmed known risk factors. BMI, body mass index; COPD, chronic obstructive pulmonary disease; NYHA, New York Heart Association.*