nature communications

Article

https://doi.org/10.1038/s41467-025-58283-7

External validation of artificial intelligence for detection of heart failure with preserved ejection fraction

Received: 4 June 2024

Accepted: 14 March 2025

Published online: 25 March 2025

Check for updates

Ashley P. Akerman[1], Nora Al-Roub[2], Constance Angell-James[2], Madeline A. Cassidy[2], Rasheed Thompson[3], Lorenzo Bosque[4], Katharine Rainer[2], William Hawkes[1], Hania Piotrowska[1], Paul Leeson[1], Gary Woodward[1], Patricia A. Pellikka[5], Ross Upton[1] & Jordan B. Strom[2]

Artificial intelligence (AI) models to identify heart failure (HF) with preserved ejection fraction (HFpEF) based on deep-learning of echocardiograms could help address under-recognition in clinical practice, but they require extensive validation, particularly in representative and complex clinical cohorts for which they could provide most value. In this study enrolling patients with HFpEF (cases; n = 240), and age, sex, and year of echocardiogram matched controls (n = 256), we compare the diagnostic performance (discrimination, calibration, classification, and clinical utility) and prognostic associations (mortality and HF hospitalization) between an updated AI HFpEF model (EchoGo Heart Failure v2) and existing clinical scores (H2FPEF and HFA-PEFF). The AI HFpEF model and H2FPEF score demonstrate similar discrimination and calibration, but classification is higher with AI than H2FPEF and HFA-PEFF, attributable to fewer intermediate scores, due to discordant multivariable inputs. The continuous AI HFpEF model output adds information beyond the H2FPEF, and integration with existing scores increases correct management decisions. Those with a diagnostic positive result from AI have a two-fold increased risk of the composite outcome. We conclude that integrating an AI HFpEF model into the existing clinical diagnostic pathway would improve identification of HFpEF in complex clinical cohorts, and patients at risk of adverse outcomes.

Heart failure (HF) is a common and morbid disease impacting an estimated 56.2 million individuals worldwide with rising prevalence[1], and impacting approximately 1 in 4 adults in their lifetime[2]. Of those suffering from HF, approximately half have a preserved left ventricular ejection fraction (HFpEF)[3–6]. HFpEF is often unrecognized or misdiagnosed due in part to insufficient consensus on its definition as well as complexities inherent in the diagnostic tools and multiple clinical pathways used for diagnosis. However, early identification and treatment may be important in limiting disease progression, reducing the burden of disease on the patient and healthcare system[7–10].

In this setting, echocardiography represents a crucial diagnostic tool for evaluating patients with undifferentiated dyspnea and suspected HFpEF due to the wealth of information it provides on

[1]Ultromics Ltd, 4630 Kingsgate, Cascade Way, Oxford Business Park South, Oxford OX4 2SU, UK. [2]Richard A. and Susan F. Smith Center for Outcomes Research in Cardiology, Division of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [3]Howard University College of Medicine, Washington, DC, USA. [4]Drexel University College of Medicine, Philadelphia, PA, USA. [5]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA. e-mail: jstrom@bidmc.harvard.edu

https://doi.org/10.1038/s41467-025-58283-7

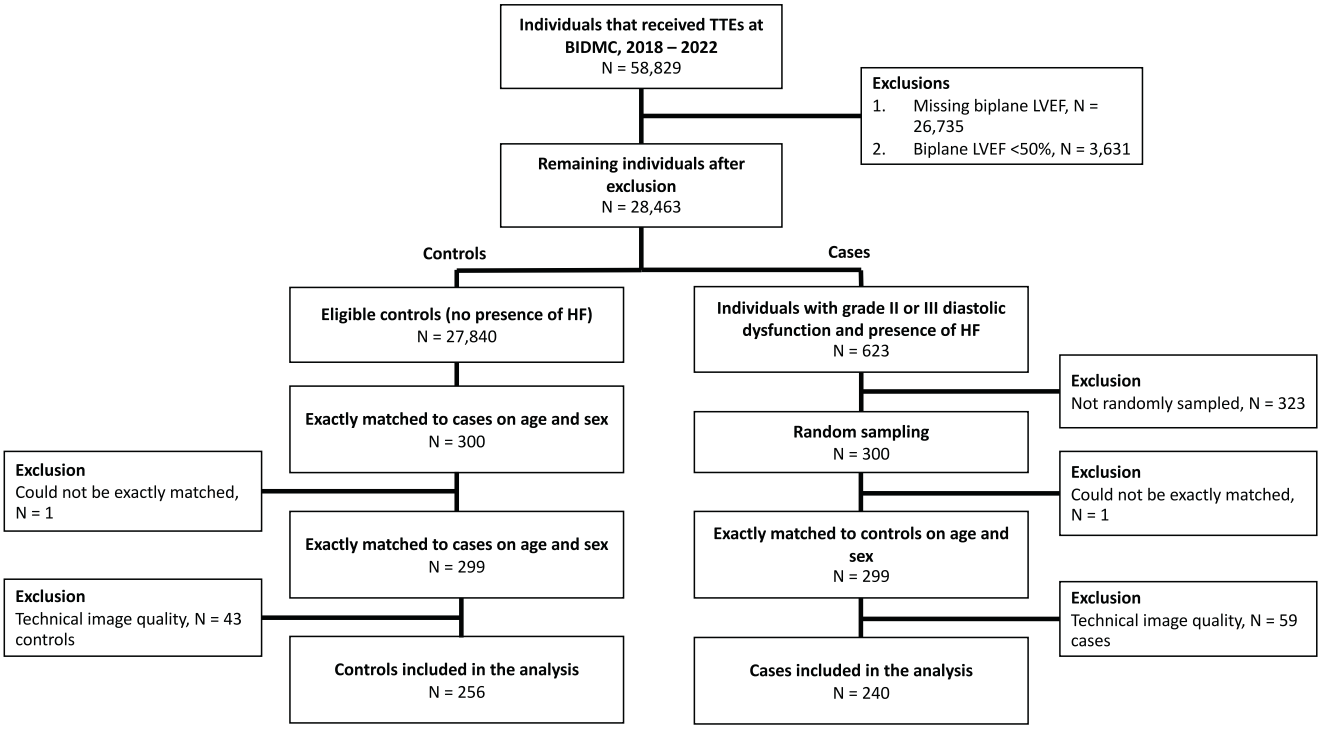

Fig. 1 | Flowchart demonstrating study inclusion and exclusions. Displayed is a flowchart depicting the study selection for cases and controls based on inclusion and exclusion criteria. BIDMC Beth Israel Deaconess Medical Center, LVEF left ventricular ejection fraction, HF heart failure, TTE transthoracic echocardiogram.

myocardial remodeling, diastolic dysfunction, estimation of left ventricular filling pressures (LVFP), chronicity of illness, and alternative causes of dyspnea such as valvular heart disease[7,11]. While recently developed multiparametric clinical scores incorporating clinical, laboratory, and echocardiographic information have shown promise in estimating the probability of HFpEF among patients with symptoms, they face limitations in cases of incomplete or inconsistent input data[12,13]. Furthermore, up to 30% of patients may be classified as "indeterminate" by echocardiographic assessment, leading to uncertainty about whom should be referred for further diagnostic testing to confirm a HFpEF diagnosis[14,15]. As such, artificial intelligence (AI) models based on deep learning of echocardiographic images have recently been developed using limited input information (a single apical 4-chamber transthoracic echocardiogram [TTE] video clip) with excellent performance in identifying those with HFpEF. While such models show promise, more validation is needed to understand how they compare to existing multiparametric clinical scores in patients at high risk for HFpEF and how AI models could be best integrated in practice to optimize patient management[16].

As such, in this study, we aimed to externally validate the performance of an AI HFpEF model (EchoGo Heart Failure, Ultromics Ltd) in an independent dataset, and compare this to existing multiparametric clinical scores for diagnosis of HFpEF in a matched case control study including patients with and without HFpEF who underwent TTE at Beth Israel Deaconess Medical Center (BIDMC). While this AI HFpEF model has previously undergone extensive multi-center internal validation across a large integrated delivery network[17], the motivation for the current study was to examine whether performance generalizes between institutions, and evaluate how architectural improvements due to experiences in clinical implementation (**Supplementary Methods**), may have impacted model performance. We hypothesized that the AI HFpEF model would outperform existing clinical scores and would be associated with patient outcomes.

Results

Patient Population

A final sample of 240 cases and 256 controls were identified after application of exclusions (Fig. 1). All cases had HF within the prior year on manual chart review and all controls lacked HF in the year prior to or following the index TTE. Overall, cases (mean ± SD; 74.2 ± 12.1 years, 54.2% female) and controls (75.0 ± 13.0 years, 55.1% female) had a similar age and sex distribution as expected due to matching (Table 1). However, cases were more likely to be Black, (18.3% vs. 6.6%, $p = 0.001$), had a higher BMI (29.4 ± 7.1 vs. 27.1 ± 5.5, $p < 0.001$) and overall had a higher number of comorbidities (Table 1). Additionally, cases had a lower estimated glomerular filtration rate, higher values for NT-proBNP, lower on-treatment cholesterol, and more frequently with cardioactive medications including beta blockers, statins, mineralocorticoid inhibitors, calcium channel blockers, antithrombotics, loop diuretics, and nitrates (all $p < 0.05$). While cases and controls had a similar biplane LVEF (63.1 ± 7.8 vs. 63.4 ± 7.0, $p = 0.67$; Supplementary Table 1), cases had a greater impairment in systolic function (global longitudinal strain -15.5 ± 4.8 vs. -20.0 ± 2.7, $p = 0.009$) and diastolic function (Supplementary Table 1; all $p < 0.05$). The proportion of cases with normal, abnormal, and missing echocardiographic and clinical parameters is summarized in Supplementary Fig. 1. However, the control group also demonstrated high rates of concentric remodeling (mean relative wall thickness, 0.48 ± 0.15), left atrial dilation, higher than normal E/e, elevated tricuspid regurgitant velocity, and elevations in NT-proBNP levels, representing a complex clinical cohort. Supplementary Tables 2–3 demonstrate that the difference between the current cohort of cases and controls, compared with those used in the development and original validation of the AI HFpEF model. Generally, the current cohort of cases and controls were the most similar (to each other) across various key clinical characteristics, than in previous datasets. Importantly, the current cohort of controls were older and more commonly had hypertension, structural heart disease, atrial fibrillation, coronary artery disease, diabetes, pulmonary

https://doi.org/10.1038/s41467-025-58283-7

Table 1 | Baseline Clinical Characteristics

Variable	Controls	Cases	*p* value
Demographics			
Age (y)	75 ± 13.0 [256]	74.2 ± 12.4 [240]	0.603
Female sex	141 (55.1%) [256]	130 (54.2%) [240]	0.910
Body mass index (kg/m^2)	27.1 ± 5.5 [252]	29.4 ± 7.1 [239]	<0.001
Race			
White	200 (78.1%)	164 (68.3%)	
Black	17 (6.6%)	44 (18.3%)	
Asian	14 (5.5%)	14 (5.8%)	0.001
Other	23 (9.0%)	17 (7.1%)	
Clinical History			
Hypertension	178 (69.5%)	209 (87.1%)	<0.001
Hyperlipidemia	149 (58.2%)	162 (67.5%)	0.039
Atrial fibrillation	48 (18.8%)	108 (45.0%)	<0.001
Permanent	10 (3.9%)	24 (10.0%)	
Paroxysmal	24 (9.4%)	42 (17.5%)	0.311
Persistent	2 (0.8%)	11 (4.6%)	
Diabetes mellitus	55 (21.5%)	100 (41.7%)	<0.001
Chronic kidney disease	37 (14.5%)	114 (47.5%)	<0.001
Coronary artery disease	50 (19.5%)	82 (34.2%)	<0.001
History of MI	25 (9.8%)	42 (17.5%)	0.016
History of CABG	22 (8.6%)	22 (9.2%)	0.922
Presence of cardiac pacemaker	19 (7.4%)	20 (8.3%)	0.832
Chronic obstructive pulmonary disease	15 (5.9%)	42 (17.5%)	<0.001
NYHA class			
I	5 (2.0%)	15 (6.3%)	<0.001
II	1 (0.4%)	57 (23.8%)	
III	0.0%	35 (14.6%)	
IV	0.0%	2 (0.9%)	
Smoking	18 (7.0%)	23 (9.6%)	0.385
History of stroke/TIA	25 (9.8%)	34 (14.2%)	0.169
Laboratory Results			
Glucose – mg/dL	114.6 ± 40.9 [224]	125.0 ± 52.1 [231]	0.018
Sodium – mEq/L	138.9 ± 9.4 [241]	138.9 ± 4.2 [239]	0.949
Potassium - mEq/L	4.3 ± 0.5 [242]	4.3 ± 0.5 [239]	0.638
Creatinine – mg/dL	1.3 ± 1.7 [243]	2.2 ± 2.2 [239]	<0.001
Estimated glomerular filtration rate – mL/min/1.73 m^2	65.7 ± 17.9 [237]	50.1 ± 24.7 [233]	<0.001
Albumin – g/dL	5.1 ± 14.0 [172]	3.8 ± 2.4 [211]	0.157
Troponin – ng/mL	0.08 ± 0.33 [130]	0.14 ± 0.58 [188]	0.286
NT-proBNP – pg/mL	2280 ± 6345 [70]	7845 ± 12820 [178]	0.001
Total cholesterol - mg/dL	172.9 ± 48.5 [137]	146.7 ± 47.0 [130]	<0.001
High Density Lipids (HDL) - mg/dL	57.3 ± 18.5 [137]	50.8 ± 19.1 [129]	0.005
Low Density Lipoprotein (LDL) - mg/dL	93.3 ± 36.6 [134]	75.4 ± 31.4 [131]	<0.001
Triglycerides - mg/dL	129.2 ± 106.0 [143]	132.4 ± 85.2 [133]	0.785
Hemoglobin – g/dL	11.7 ± 2.3 [234]	9.9 ± 2.4 [238]	<0.001
Treatment			
Beta-blocker	105 (41.0%)	174 (72.5%)	<0.001
Aspirin	99 (38.7%)	112 (46.7%)	0.101
Statin	133 (52.0%)	165 (68.8%)	<0.001
Other lipid lowering drug	11 (4.3%)	16 (6.7%)	0.333
Mineralocorticoid antagonist	10 (3.9%)	23 (9.6%)	0.020
SGLT2 inhibitor	2 (0.8%)	1 (0.4%)	>0.999
ACE inhibitor	63 (24.6%)	43 (17.9%)	0.076
Angiotensin receptor blocker	43 (16.8%)	54 (22.5%)	0.148
Calcium channel blocker	66 (25.8%)	104 (43.3%)	<0.001
Antithrombotic	44 (17.2%)	99 (41.3%)	<0.001

Table 1 (continued) | Baseline Clinical Characteristics

Variable	Controls	Cases	*p* value
Loop diuretic	36 (14.1%)	158 (65.8%)	<0.001
Anti-platelet agent	28 (10.9%)	28 (11.7%)	0.932
Nitrate	13 (5.1%)	41 (17.1%)	<0.001
Antiarrhythmic medication	11 (4.3%)	17 (7.1%)	0.259
Gout medication	28 (10.9%)	32 (13.3%)	0.515

Shown are the clinical characteristics of the study population, stratified by cases and controls. P values for the comparison of cases vs. controls are shown using two-sided Fisher's exact tests for categorical variables and Student's t tests for continuous variables. Results are shown as mean ± standard deviation and effective sample size in [square brackets], and count (proportion), where appropriate. Comparisons of baseline echocardiographic variables are provided in Supplementary Table 1.
ACE angiotensin-converting-enzyme, *CABG* coronary artery bypass graft, *GFR* glomerular filtration rate, *MI* myocardial infarction, *NT-BNP* N-terminal B-type natriuretic peptide, *NYHA* New York Heart Association, *SGLT2* sodium/glucose cotransporter 2, *TIA* transient ischemic attack.

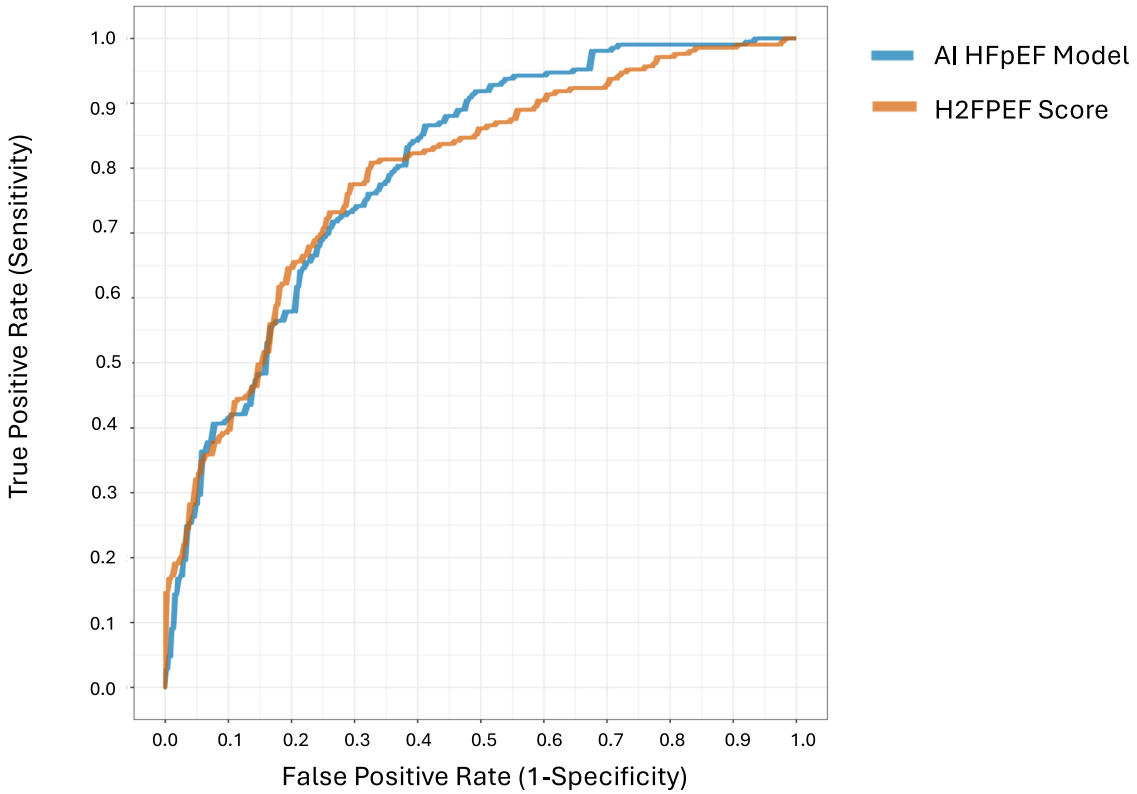

Fig. 2 | Receiver operating characteristic curve comparing discrimination for identification of heart failure with preserved ejection fraction (HFpEF) using artificial intelligence (AI) HFpEF model versus the H2FPEF score. Shown are the receiver operating characteristic (ROC) curves comparing discrimination for identification of HFpEF using AI HFpEF model vs. the H2FPEF score. AI HFpEF is in blue (area under the curve of the ROC [AUROC]: 0.798, [95% CI 0.756–0.799]), and the H2FPEF score is in orange (0.788, ([0.745–0.789])). The difference between the two was not significant (mean difference in AUROC, 0.01, [−0.043–0.064], p = 0.710 using a two-sided DeLong test).

disease, previous cardio- and cerebro-vascular events, and HFpEF "mimics" than in previous datasets, supporting the notion of a far more complex clinical cohort.

Model discrimination and calibration

Discrimination was similar between the AI HFpEF model (AUROC: 0.798, 95% CI 0.756–0.799) and H2FPEF score (AUROC: 0.788, 0.745–0.798, difference: p = 0.71; Fig. 2). Both the AI HFpEF model and H2FPEF score demonstrated similar calibration results. The intercept and slope for the AI HFpEF model (intercept: −0.56, −0.82 to −0.32; slope: 0.44, 0.35 to 0.56) and H2FPEF score (intercept: −1.00, 1.75 to −0.40; slope: 0.81, 0.58–1.15) both indicated an overestimation of risk. For the AI HFpEF model this overestimation was predominant in moderate-to-high probabilities, whereas for the H2FPEF score there was overestimation across the full spectrum of probabilities (Supplementary Fig. 3, Hosmer-Lemeshow, both p < 0.001).

Model classification

Table 2 highlights classification performance for all models. Full confusion matrices are provided in Supplementary Table 4, and comparison of performance across sociodemographic groups, and relevant clinical risk factors in Supplementary Tables 5–10. All three clinical scores provided a "intermediate" classification when there was discordant, or missing information to reliably suggest HFpEF presence or absence. Classification performance is therefore presented when considering all patients (i.e., including intermediate outputs in the calculation of classification statistics; Supplementary Tables 11–12), and when considered only diagnostic outputs (i.e., positive or negative; **Supplementary Methods** and Supplementary Tables 11–12).

Table 2 | Classification of heart failure with preserved ejection fraction (HFpEF) according to three models

Model	Calculation Method	Sensitivity	Specificity	NPV	PPV
AI HFpEF	*All Data*	77.4% (74.6–79.6%)	50.2% (48.6–52.1%)	81.6% (73.6%–87.6%)	67.3% (59.6%–71.7%)
	Diagnostic Only	86.6% (80.2–90.9%)	58.5% (51.0–64.8%)	81.6% (73.6%–87.6%)	67.3% (59.6%–71.7%)
	Ignore Uncertainty	83.8% (78.2–88.1%)	53.9% (47.5–60.0%)	78.0% (70.3–82.8%)	63.0% (56.1–67.0%)
H2FPEF	*All Data*	53.9% (50.2–58.2%)	12.8% (11.8–13.9%)	90.3% (75.0–100.0%)	73.6% (64.9–79.3%)
	Diagnostic Only	97.5% (94.0–100.0%)	40.0% (26.8–52.9%)	90.3% (75.0–100.0%)	73.6% (64.9–79.3%)
HFA-PEFF	*All Data*	63.2% (58.3–68.0%)	29.0% (26.9–31.0%)	98.5% (94.7–100.0%)	86.3% (79.3–91.1%)
	Diagnostic Only	99.3% (97.3–100.0%)	74.7% (62.8–83.8%)	98.5% (94.7–100.0%)	86.3% (79.3–91.1%)

Classification statistics and 95% confidence intervals (CIs) are provided for the three models using all available data (i.e., including those with intermediate classifications in the calculation), and when only considering diagnostic studies (i.e., **excluding** those with intermediate classifications in the calculation). For the AI HFpEF model, classification performance is also presented when uncertainty metrics are disregarded, and classification of patients is based only on prediction class probability (i.e., ≥0.50). For further information related to the calculation of these statistics, please see Supplementary Table 12.

AI HFpEF artificial intelligence model for the prediction of heart failure with preserved ejection fraction, *HF2PEF* Heavy, Hypertensive, Atrial Fibrillation, Pulmonary Hypertension, Elderly, Filling Pressure, *HFA-PEFF* Heart Failure Association Pre-test assessment, Echocardiography and Natriuretic Peptide Score, Functional Testing in Case of Uncertainty, Final Etiology, *NPV* negative predictive value, *PPV* positive predictive value.

H2FPEF score

The H2FPEF score resulted in 306 (61.7%) intermediate classifications, compared with 75 (15.1%) for the AI HFpEF model. Compared to the H2FPEF score, the AI HFpEF model demonstrated higher sensitivity (mean difference, 26.7%, 17.9–35.4%, $p < 0.001$), and specificity (mean difference, 37.5%, 29.9–45.1%, $p < 0.001$), but no difference in NPV (−8.7%, −22.8 to 5.3%, $p = 0.36$), and PPV (−6.3%, −15.7 to 3.1%, $p = 0.21$). When considering only diagnostic outputs (i.e., removing "intermediate" scores from the calculation), the H2FPEF score demonstrated higher sensitivity ($p = 0.002$) and specificity ($p = 0.011$).

Owing to the meaningful impact of intermediate classifications on model performance, due either to missing or discordant clinical parameters (detailed in **Supplementary Methods**), classification was also assessed only in patients with all information available (i.e., true discordance contributing to intermediate classifications, rather than missing data). After removing patients with missing inputs for the H2FPEF score (n = 319 [64.3%]), H2FPEF sensitivity improved (97.5%, 92.5–100%, vs. 86.3%, 79.8–91.2% for the AI HFpEF model), and specificity improved (38.6%, 18.7-53.1% vs. 59.9%, 49.2–70.7% in the AI HFpEF model), despite retaining high intermediate classification rates (61.1% vs. 13.5% in AI HFpEF model; Supplementary Table 11).

Supplementary Table 13 illustrates how performance of the H2FPEF changes as a consequence of the decision threshold. As expected, higher decision thresholds result in lower sensitivity, and higher specificity, and the converse is true for lower decision thresholds. Sensitivity ranged from 3.8% to 98.8%, and specificity ranged from 10.9% to 100%, for thresholds of categorical scores (1–9). Supplementary Table 6 highlights how the AI HFpEF model might compensate for the drop in performance above a score of 6, and below a score of 4 (sensitivity and specificity respectively). Similarly, Supplementary Table 14 demonstrates consistent performance between the AI HFpEF model and the H2FPEF score for comparable decision thresholds according to the continuous probability produced by each model. Additionally, Supplementary Tables 5–10 demonstrate consistent model performance across relevant sociodemographic, clinical, and echocardiographic subgroups of interest.

HFA-PEFF Score

There were 269 (54.2%) intermediate classifications with the HFA-PEFF score. Compared with the HFA-PEFF score, the AI HFpEF model also demonstrated higher sensitivity (by 17.5%, 8.8 to 26.2%, $p < 0.001$), specificity (by 23.0%, 14.5–31.6%, $p < 0.001$), but lower NPV (by −16.9%, −9.0 to −24.8%, $p = 0.002$), and PPV (by −19.0%, −10.8 to −27.3%, $p < 0.001$). However, when considering only diagnostic outputs, the HFA-PEFF score also demonstrated higher sensitivity ($p < 0.001$) and specificity ($p = 0.012$). Supplementary Table 13 illustrates how performance of the HFA-PEFF score changes as a consequence of the decision threshold, demonstrating similar performance characteristics as the H2FPEF score. After removing patients with missing inputs for the HFA-PEFF score (n = 246 [49.6%]), sensitivity improved to 100% vs. 87.0% (81.4-92.1%) in AI HFpEF model, but specificity decreased (26.7%, 9.8-42.2% vs. 52.6%, 37.3–69.1% for AI HFpEF model), and intermediate classification rates also decreased (31.3% vs. 14.2% in AI HFpEF model; Supplementary Table 11).

Reclassification and added information

Figure 3 illustrates the reclassification of patients from the H2FPEF Score (panel A) and HFA-PEFF Score (panel B) into the AI HFpEF model, alongside net classification statistics for the AI HFpEF model compared to each score. In the intermediate classified patients according to the H2FPEF score (n = 306), the AI HFpEF model also classified 45 as intermediate, 163 as high likelihood of HFpEF, and 98 as low likelihood of HFpEF. In the intermediate classified patients according to the HFA-PEFF score (n = 269), the AI HFpEF model also classified 48 as intermediate, 139 as high likelihood of HFpEF, and 82 as low likelihood of HFpEF.

For the continuous probability output from the AI HFpEF model and the H2FPEF score, the AI HFpEF model provided significantly more information (likelihood ratio test, $p < 0.001$), supported by positive overall net reclassification statistics (NRI 0.40, 0.21-0.59; Supplementary Table 15), and marginally positive discrimination improvement (IDI) index (0.07, 0.00-0.14). However, when utilizing only the categorical outputs from the three models, the AI HFpEF model added significant information to both the H2FPEF and HFA-PEFF scores according to likelihood ratio tests (both $p < 0.001$), but NRI statistics were less consistent (Fig. 3 and Supplementary Table 15).

Clinical utility

The clinical utility of integrating the AI HFpEF model into clinical practice was assessed using decision curve analysis[18]. Making clinical management decisions based on the integration of diagnostic information from the H2FPEF score and the AI HFpEF model was superior to the H2FPEF score alone. Specifically, compared to only using the H2FPEF score, decisions to intervene using all available information resulted to 33% more patients with HFpEF managed correctly, and an absolute 9% reduction in the number of prescriptions without missing any patients with HFpEF; this benefit was highest when clinical scores were used following intermediate AI HFpEF model classifications, rather than the alternative (Fig. 4 and Supplementary Table 16). Comparisons to the HFA-PEFF score demonstrated little difference between approaches (Supplementary Fig. 4). Modeling of combined continuous probabilities (AI HFpEF model and H2FPEF Score) and categorical classification (HFA-PEFF Score) supported the notion of

https://doi.org/10.1038/s41467-025-58283-7

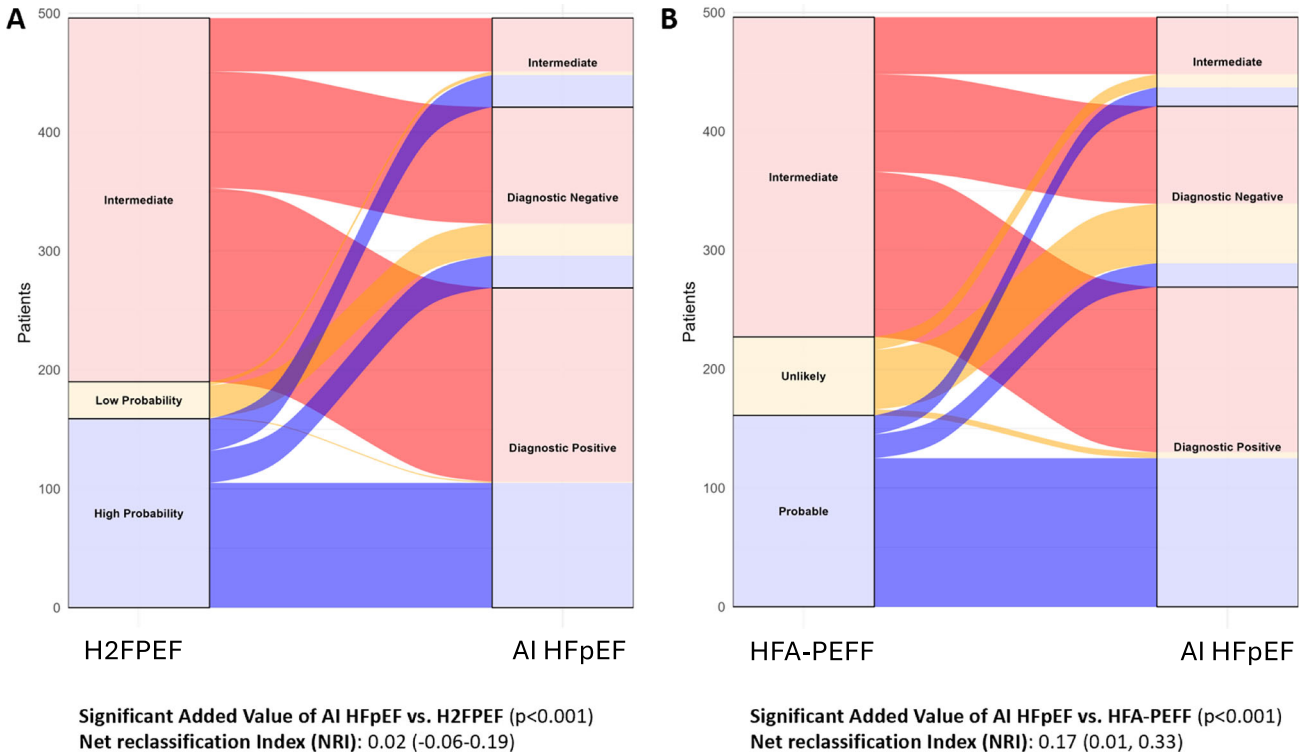

Fig. 3 | Alluvial plots demonstrating reclassification of predicted heart failure with preserved ejection fraction (HFpEF) from clinical scores using the artificial intelligence (AI) HFpEF model. Displayed are alluvial plots depicting the reclassification of predicted HFpEF from clinical scores using the AI HFpEF model. This plot and associated reclassification statistics account for only categorical classification outputs from each model, rather than continuous outputs. Panel A depicts the AI model's reclassification of an individual's predicted HFpEF status from the HF2FPEF score. Panel B depicts the AI model's reclassification of an individual's predicted HFpEF status from the HFA-PEFF score. The added value of the AI HFpEF model compared with the H2FPEF score and HFA-PEFF score are displayed below the alluvial plots. Two-sided likelihood ratio tests were used to estimate the added value of the AI HFpEF vs. H2FpEF score and resulting p values are presented alongside net reclassification improvement (NRI) statistics. All NRI statistics are based on categorical outputs. Non-diagnostic and indeterminate outputs are referred to as "intermediate" for consistency and clarity.

added clinical utility when integrating information from all available diagnostic methods (Supplementary Table 16).

Patient outcomes

During a median (IQR) 25 (15–35) months follow-up, there were 45 HF hospitalizations (10.3%), and 61 deaths (14.2%). A diagnostic positive result according to the AI HFpEF model was associated with a two-fold risk of the primary composite endpoint (Fig. 5, Supplementary Table 17, HR 2.56, 1.46–4.51, p = 0.001 vs. diagnostic negative output), risk for mortality (HR 2.54, 1.27–5.05, p = 0.008), and three-fold higher risk of HF hospitalization (HR 3.15, 1.33–7.47, p = 0.009). The same association was present for quartiles of risk according to the continuous probability of HFpEF from the AI HFpEF model. Compared to the first quartile, risk of the composite outcome (HR 3.95, 2.00–7.81, p < 0.001), all-cause mortality (HR 3.14, 1.40–7.02, p = 0.005), and HF hospitalization (HR 4.74, 1.61–13.9, p = 0.005) were all elevated (Supplementary Table 18). Similar associations were also present for the clinical scores (Supplementary Table 17; Supplementary Figs. 5–6), but there was no significant association between the AI HFpEF model and risk of outcomes in intermediate classified patients according to the H2FPEF and HFA-PEFF scores (Supplementary Table 19).

Discussion

In this retrospective case-control study, an AI HFpEF model using deep learning of a single TTE video clip had higher classification performance than existing clinical scores, largely owing to fewer intermediate classifications. When clinical scores provided a diagnostic output, they demonstrated excellent classification performance. The continuous output of the AI HFpEF model demonstrated a significant increase in the information available for diagnosis, beyond that of the clinical scores, but this was less clear for categorical outputs. Overall, combining use of AI and clinical score information resulted in greater rates of appropriate treatment for HFpEF and fewer unnecessary treatments in decision modeling. All algorithms, including both the AI HFpEF model and clinical scores, demonstrated significant associations with risk of all-cause mortality and HF hospitalization. Overall, these results indicate a possible role for this AI HFpEF model in the HFpEF diagnostic pathway, particularly in combination with information from clinical scores, to identify patients warranting confirmatory testing or management for HFpEF.

While HFpEF is highly prevalent, impacting up to 64 million individuals worldwide with 5.7 million in the US alone and associated with a 5-year survival as low as 35%, it can be challenging to make the diagnosis currently due to (1) absence of a clear standardized definition, (2) complexity of existing echocardiography-based diastology criteria, and (3) heterogeneity of HFpEF etiology[8,19]. While TTE remains key to establishing the presence of increased LVFP and diastolic grade, individual TTE parameters are only modestly related to LVFP[11,20–25]. Moreover, while guideline-based approaches have been developed to overcome this deficiency through a multiparametric approach to incorporating TTE findings, up to 30% of individuals may have indeterminate/ intermediate results from these algorithms due in part to missing or discordant TTE parameters, resulting in poor sensitivity for detection of HFpEF, particularly early in the disease state[11,14,15,25]. In response to such challenges with using existing TTE variables for diagnosis, multiparametric clinical scores have been developed

https://doi.org/10.1038/s41467-025-58283-7

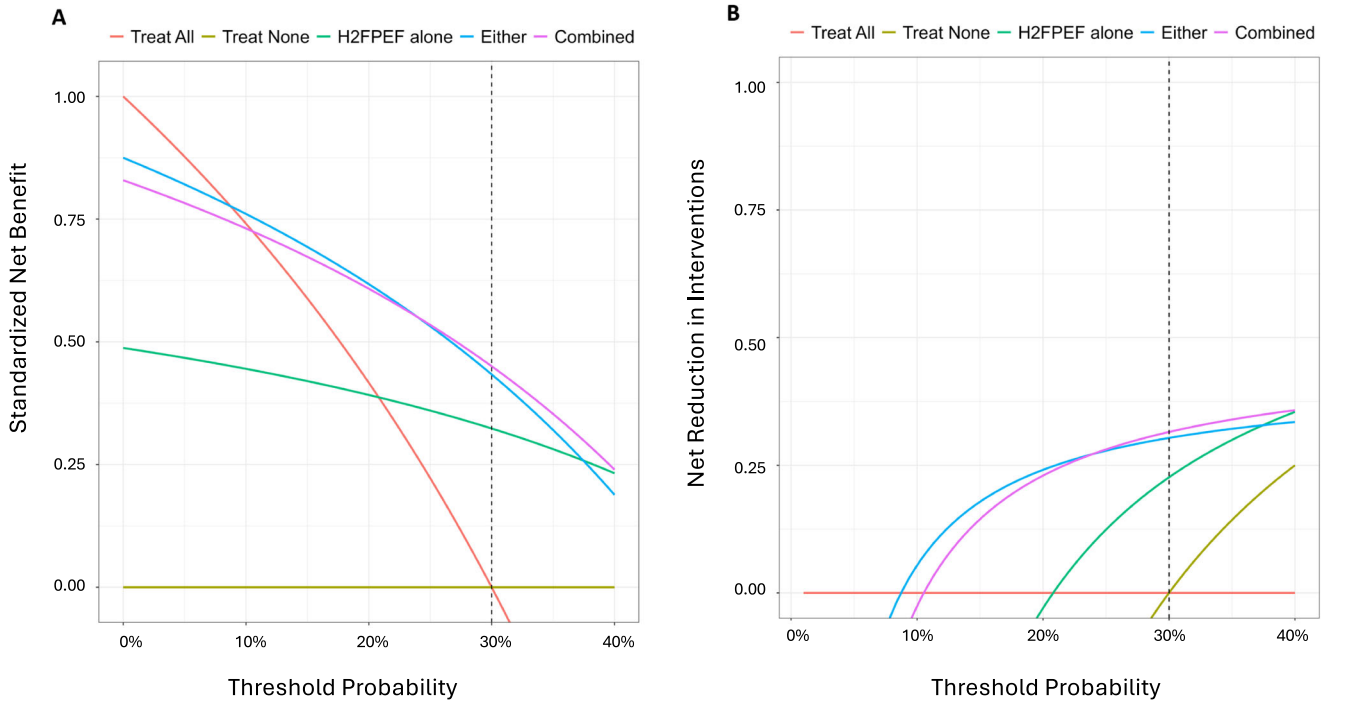

Fig. 4 | Decision curves demonstrating standardized net benefit and net reduction in interventions of using the H2FPEF and AI HFpEF model in combination versus separate approaches in patients suspected of having HFpEF. Decision curves comparing the standardized net benefit (panel **A**) and net reduction in interventions (**B**) when patient management decisions are based on the output of the H2FPEF score and/or the AI HFpEF model. The management decision is assumed to represent prescription of SGLT2i to the patient, in a population where the expected prevalence of HFpEF is 30%. Red and gold lines represent clinical baselines of prescribing all patients or no patients SGLT2i, respectively, regardless of the output of any test. Prescribing based on only the H2FPEF score (green), represents that any patient with a "Probable" classification of HFpEF would be prescribed SGLT2i. Prescribing based on either a "Probable" (H2FPEF) or "Positive" classification (AI HFpEF) is presented in blue. Prescribing based on the combination of a "Positive" classification (AI HFpEF), or "Intermediate" (AI HFpEF) and "Probable" (H2FPEF) is presented in purple. The x-axis represents the threshold probability that would be required by a clinician and/or patient to initiate prescription of SGLT2i. In this context, the chosen minimum threshold probability is 30% (dotted line), representing the relative harm of an adverse event when taking SGLT2i (5.8%[41]), and the risk reduction associated with taking an SGLT2i (−19%[41] for HF hospitalization or worsening HF event). The x-axis is truncated to clinically reasonable threshold probabilities for clear and meaningful interpretation. For net benefit plots, the y-axis refers to the standardized net benefit of taking a given approach, with units presenting the proportions of patients with disease in the population who would be successfully managed according to the different approaches. For example, a value of 0.45 for the Combined approach would be interpreted such that, compared to prescribing no patients with SGLT2i, managing patients based on the combined information from the AI HFpEF model and the H2FPEF score would result in 45% of patients with HFpEF being correctly managed. For net reduction in interventions, a value of 0.315 for the Combined approach would be interpreted such that, prescribing SGLT2i based on the combined information would lead to an absolute 31.5% reduction in the number of prescriptions without missing any patients with HFpEF.

incorporating clinical, echocardiographic, and biomarker information to screen patients for HFpEF, though these too may be inconclusive in the setting of missing or discordant clinical data, and are complex to integrate into practice, requiring significant a priori knowledge of a patient's clinical history and referral for laboratory and echocardiographic testing[12,13]. As such, at present, nearly 45% of patients with early HFpEF present with normal LVFP at rest and require invasive testing or stress echocardiography to demonstrate impairments[12].

In this setting, there has been great enthusiasm in the use of AI computer vision technologies to improve early diagnosis and treatment of HFpEF to delay progression and avoid hospitalization[26]. One such AI algorithm, the Ultromics EchoGo Heart Failure model, applied to a single 4-chamber TTE video clip, has shown to have a high sensitivity (87.8%) and specificity (81.9%) in independent testing to determine an individual's probability of HFpEF[17]. In the current study, applying an updated version model to an independent validation cohort, the AI HFpEF model had similar performance to multiparametric clinical scores to identify HFpEF, with an AUROC of 0.798 (95% CI: 0.756–0.799) compared to 0.788 (0.745-0.798) for the H2FPEF score, but resulted in fewer intermediate classifications, with only 15.1% having a non-diagnostic classifications compared to 61.7% and 54.2% of those classified according to the H2FPEF and HFA-PEFF scores, respectively. Consequently, classification performance was higher using the AI HFpEF model than clinical scores when considering these intermediate classifications, but lower when patients with intermediate classifications were removed. This supports previous investigations[12,13] demonstrating that existing clinical scores have excellent performance (particularly sensitivity) when they provide a clear diagnostic output, but are limited by high intermediate classification rates, which might be more apparent in situations of low data quality (e.g., missing, poor imaging quality) or true clinical uncertainty (discordant parameters). Sub-groups analysis of patients in whom the multiparametric models were all available (i.e., retaining only clinical discordance), demonstrated improved clinical score performance, but retained high rates of intermediate classifications (61% for H2FPEF, 52% for HFA-PEFF), indicating that clinical discordance remains a critical issue during the diagnostic pathway.

Discrimination performance was overall worse than on internal validation (e.g., AUROC 0.798 vs. 0.95)[17]. This degree of performance decrement may be expected when applying an optimistic model to an external cohort[27]. Nevertheless, this underlines the need for external validation to identify how such models generalize to real-world clinical practice, particularly as they are further developed[28]. Changes to this AI HFpEF model following experience in real-world implementation have yet to be examined[29], and as such, this study therefore represents initial validation experience in a different integrated healthcare

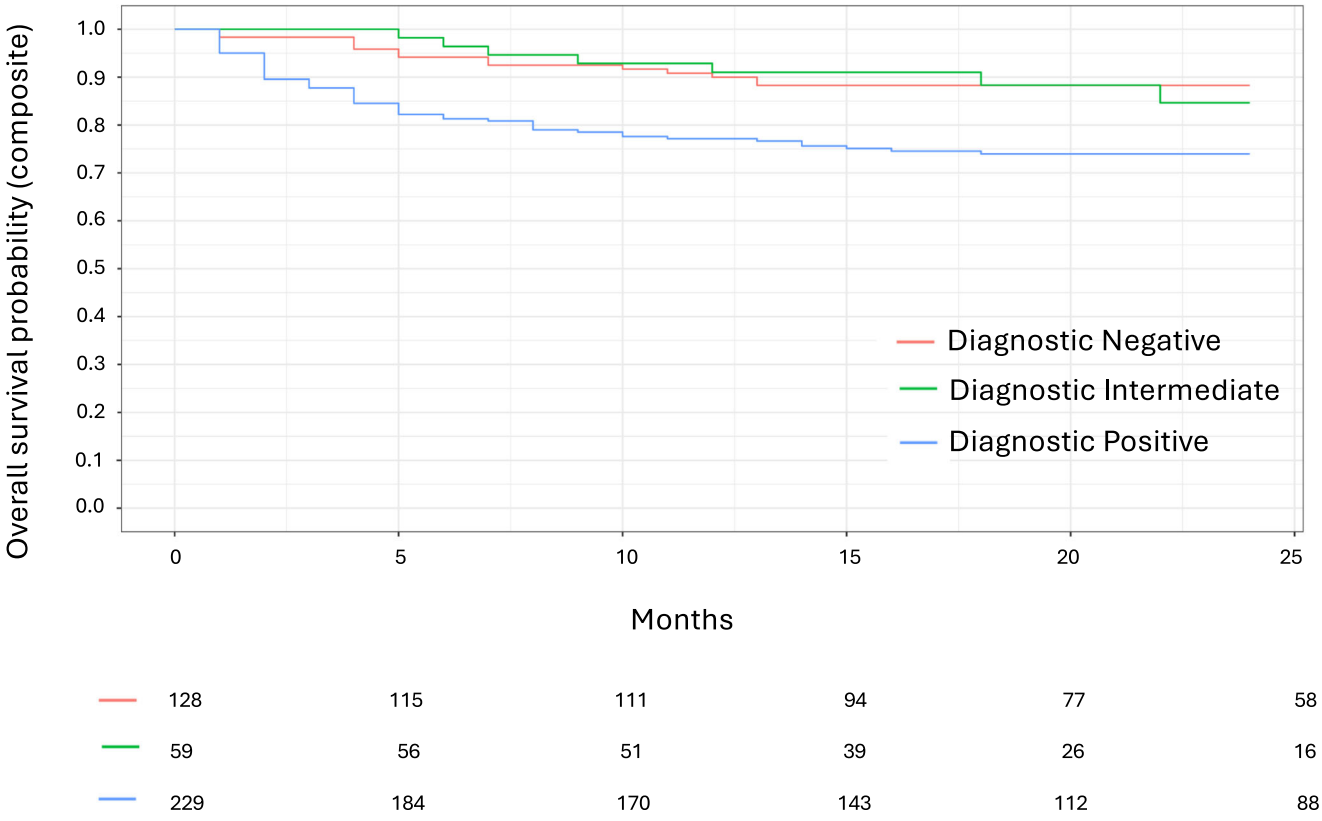

Fig. 5 | Kaplan-Meier curve demonstrating time to the composite endpoint by predicted classification according to the artificial intelligence (AI) heart failure with preserved ejection fraction (HFpEF) model. Shown are Kaplan-Meier curves for time (in months) from the index echocardiogram to the composite outcome of death or heart failure hospitalization according to the AI HFpEF model's predicted classification. Red = diagnostic negative, green = intermediate ("non-diagnostic" due to high uncertainty), and blue = diagnostic positive. Number of individuals in the risk set at 5-month time intervals is provided below the x-axis.

delivery network (Mayo Clinic vs. Beth Israel Lahey Health) where care patterns may differ. Despite the observed drop in discrimination, the AI HFpEF model and existing clinical scores demonstrated similar classification characteristics as those previously observed[17]. However, this is not the case for specificity, which was lower than previously reported due to the diagnostic challenge (to all models) facilitated by a complex clinical cohort (including HFpEF mimics).Importantly, the integration of the AI HFpEF model demonstrated similar utility and associations with clinical outcomes, highlighting that such iterative improvements to the model appear to have retained overall diagnostic and prognostic performance and generalizability.

Suggestions to integrate AI models into existing diagnostic pathways[16,17] and utilize the combined clinical information from all available sources are supported by the current study. In decision modeling, use of combined information from the AI HFpEF model and clinical scores resulted in a greater number of individuals appropriately receiving treatment (in this case SGLT2 inhibitors) with fewer number of individuals treated unnecessarily. This is logical, given that each model is associated with its own development process (AI, statistical modeling, and clinical consensus for AI HFpEF, H2FPEF, and HFA-PEFF, respectively). In a heterogeneous syndrome like HFpEF, each model would likely capture different types of patients under the broad umbrella of HFpEF. This feature becomes especially important given the absence of one single standard definition of HFpEF, and heterogeneity of the etiology of the syndrome; different diagnostic models which can be readily integrated would capitalize on such ambiguity and provide the greatest value for patients. The continuous output of the AI HFpEF model added significant value to the information contained in the existing clinical scores, which is likely due to the improved AI HFpEF model calibration (beyond that apparent in version 1). However, whether the continuous or categorical output would be used in clinical practice has yet to be prospectively evaluated, and exactly how different models and associated decision thresholds may be combined, and in what type of patient(s), requires extensive validation. This is important, particularly considering that the greatest benefit was observed when all models were utilized for identification of disease and decision-making processes. This approach is supported by findings in other disease states such as prediction of atrial fibrillation[30] and coronary artery disease risk[31], where the use of AI has been found to be complementary to clinical information. However, the implementation of AI in a supportive role for clinical diagnostics is still in its infancy, and requires prospective testing to understand where and how this integration might provide value.

In practice, patients with undifferentiated dyspnea could be screened using a clinical score or AI model and, if inconclusive/intermediate, alternative models could be applied to identify those that are truly at lower risk versus those with more uncertainty for which confirmatory testing (e.g. stress testing or invasive testing with right heart catheterization) could be performed. While this strategy requires prospective testing to evaluate its impact on care patterns and outcomes, results from this validation study suggest promise of a combined clinical and AI approach to improve upon existing HFpEF risk stratification, and hopefully reduce the burden of under-recognition in clinical practice[32]. A possible strategy for future deployment and implementation might require an evaluation of the decision making strategy across multiple clinical models, such that each model capitalizes on its strengths, and supports any weaknesses of others. Though this remains speculative at present, as the AI HFpEF model requires only a single 4-chamber video clip, it is possible that future iterations of the model applied to point-of-care ultrasound (POCUS)

images could permit risk stratification at the point of care and rapid triage and treatment with the goal of slowing disease progression, providing clarity around the HFpEF diagnosis, and avoiding hospitalizations.

While including a rigorously phenotyped population with and without HFpEF, the current study has limitations worth acknowledgement. First, it must be acknowledged that the clinical HFpEF syndrome has no clear and consistent definition, and a heterogenous etiology, meaning that not all HFpEF phenotypes and definitions will be captured in the current study. While the definitions adopted herein are consistent with recent HF guidelines[33], other patients who might be reasonably associated with the HFpEF syndrome might not be adopted under the current definition; (i) patients in whom current (or prior) ejection fraction values are below 50%, (ii) left ventricular filling pressure confirmed via alternative methods such as exercise echocardiography or right heart catheterization, and (iii) patients who might have been hospitalized due to HF at different clinical sites, or captured at different phases of the diagnostic pathway (e.g., POCUS or advanced HF clinics). Second, categorical outputs from all three models incorporate a non-diagnostic, or "intermediate" classification intended to support further confirmatory testing; direct comparison between models therefore must consider whether such intermediate classifications are due to missingness or discordance, whether this would occur in clinical implementation, and the potential diagnostic information they might contribute. For instance, NT-proBNP was missing in nearly a quarter of patients, reflecting the challenges of guideline adherence in clinical practice as well as the clinical scores that utilize such laboratory markers, but could be obtained as required. When intermediate classifications were considered a relevant element in the decision-making process, the current study demonstrated superiority of the AI HFpEF model. However, it must also be acknowledged that several patients were excluded due to poor image quality. While this degree of poor image quality is within published norms for echocardiography[34], it nevertheless exerts a greater effect on the AI HFpEF model compared to multiparametric clinical scores. Third, the continuous outcome probability of HFpEF from the AI HFpEF model, similar to the raw score values for the H2FPEF score, may provide additional information on risk beyond the dichotomized results. Finally, further training and development are likely required to ensure that the intended use population(s) are appropriately accounted for, and model improvements such as calibration in mid-range probabilities are clear areas where improvement would benefit clinical implementation (e.g., in uncertain populations); extensive retrospective and prospective validation will be required to ensure that such features provide the intended benefit to clinicians and patients.

In this case control study, an AI HFpEF model using deep learning of a single TTE video clip had higher classification performance than existing clinical scores, largely owing to high rates of intermediate classifications from multivariable scores comprised of discordant clinical parameters. The continuous output of the AI HFpEF model demonstrated added value to the diagnostic process, but the greatest benefit was observed when information from existing clinical scores and AI are integrated into the decision-making process. These results overall suggest a possible role for a combined clinical and AI approach towards the recognition of HFpEF, ultimately with the goal of reducing uncertainty in HFpEF diagnosis and ensuring timely and appropriate treatment for this high-risk population.

Methods

Ethical declaration

The research presented in this manuscript was reviewed and approved by the Institutional Review Board at BIDMC (FWA00003245) which issued a waiver of informed consent given the retrospective nature of the study and lack of feasibility to recontact all patients, several of whom have died. Several authors (N.A., C.A.J., R.T., L.B., K.R.) were involved in data collection which required use of patient identifiers (i.e. medical record number and information on TTE and outcome dates), but this information was stored at BIDMC and transmitted to Ultromics for analysis using unique study identifiers such that no protected health information was shared externally.

Study population

Individuals who received a TTE between 2018 and 2022 at BIDMC, a large tertiary care hospital of Harvard Medical School in Boston, Massachusetts, were considered for inclusion in a retrospective case-control study. This time window was chosen to reflect contemporary management of HFpEF and permit abstraction of structured TTE data on biplane left ventricular ejection fraction (LVEF), not available prior to 2018. Only an individual's first TTE during the study window was considered. All patients were required to have a biplane LVEF ≥ 50% on the index TTE and have technically evaluable images. The focus on biplane LVEF was for several reasons: (1) there may be possible differences across LVEF techniques[35], (2) current American Society of Echocardiography guidelines advocate for use of biplane LVEF[35], and (3) the criteria for ejection fraction in the development of the AI HFpEF model used biplane LVEF values[17]. The study was not pre-registered, and no protocol was previously published. The study was conducted and reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD-AI) guidance[36] (Supplementary Table 20).

Identification of cases and controls

Cases were identified through query of the institutional TTE database as having grade II-III diastolic dysfunction. Additionally, cases were required to have evidence of recent hospital admission for heart failure (HF) within a year prior to the index TTE date, as evidenced by a primary discharge diagnosis of 428.X (under the International Classification of Diseases, 9th Revision [ICD-9] classification) prior to October 1, 2015 or I50.X (under the International Classification of Diseases, 10th Revision [ICD-10] classification) on or after October 1, 2015. Controls were similarly identified as having grade I, or no diastolic dysfunction on TTE, and no evidence of a hospital admission for HF within one-year prior to and after the index TTE date. Presence or absence of HF hospitalization was additionally adjudicated and confirmed at the time of manual chart review, but this information was not used in case or control selection. Among possible cases and controls, 300 cases were randomly selected and exactly matched on age, sex, and year of index TTE to control subjects (Fig. 1).

Patient characteristics and outcomes

Detailed medical chart review of included patients was performed by trained analysts (NA, CAJ, RT, LB, KR). Study data were collected and managed using Castor Electronic Data Capture, a secure, web-based electronic data capture tool[37,38]. A detailed list of all demographic, clinical, echocardiographic, and treatment variables and their definitions is provided in Supplementary Table 21. Echocardiographic variables were abstracted from TTE reports without re-review of images. Biplane left ventricular ejection fraction values were not verified by manual inspection of images, but instead, values from the final study report (by experienced faculty echocardiographers) were extracted, and thus reflect the subsequent patient management pathway. Outcomes included all-cause mortality and HF hospitalization within the year following the index TTE, with occurrence of an outcome abstracted by medical chart review. Patient status (case or control) was not blinded to analysts during extraction of relevant clinical parameters and patient outcomes.

AI HFpEF model

The development and initial validation of the AI HFpEF model (EchoGo Heart Failure v2, Ultromics Ltd, United Kingdom) has been previously

described in depth[17]. Briefly, a three-dimensional (3D) convolutional neural network (CNN) was developed on apical four chamber images from TTE videos. All patients had undergone comprehensive TTE at Mayo Clinic (Rochester, United States of America) or St. Georges Hospital (National Health Service, United Kingdom). For the derivation and initial validation study, cases were identified according to AHA/ACC/HFSA guidelines[33], as having (i) presence of a relevant ICD9/10 code for HF within one year of the TTE; (ii) preserved systolic function evidence by an LVEF ³ 50%; and (iii) documented presence of increased intracardiac LVFP on TTE and were matched on age, sex, and year of TTE to the controls, similar to the design of the current study[33]. Multicenter external validation was conducted at clinical sites within the Mayo Clinic Health System, spanning 4 states, using up-sampling of non-White and Hispanic populations to improve generalizability of results.

The AI HFpEF model provides a continuous prediction class probability value between 0 and 1, which was mapped to a binary negative and positive diagnostic prediction of HFpEF, respectively. The classification threshold for the output predictions on the validation dataset was set to 0.5. A third categorical output, "no classification", herein referred to as "intermediate," is generated based on model instability and uncertainty, to ensure that unstable and uncertain predictions are not provided to a user (**Supplementary Methods**). The term intermediate is used to highlight that important diagnostic information can still be contained within uncertain predictions. The same terminology is also used for the multiparametric clinical scores described below. In the initial derivation study which included 2,971 cases and 3,785 controls, version 1 of the AI HFpEF model demonstrated excellent discrimination for both training (area under the receiver operating characteristic curve [AUROC], 0.97, 95% CI 0.96–0.97) and validation (0.95, 0.93–0.96) data, which was maintained in independent testing (0.91, 0.90–0.93), and identified individuals at greater risk of mortality[17]. Changes made in version 2 are described in detail in the **Supplementary Methods**.

Multiparametric clinical scores

AI HFpEF model results were compared with two widely used and previously validated multiparametric clinical scores, the H2FPEF and HFA-PEFF scores[13]. The H2FPEF score (**H**eavy, **H**ypertensive, Atrial **F**ibrillation, **P**ulmonary Hypertension, **E**lderly, **F**illing Pressure) and HFA-PEFF score (**H**eart **F**ailure **A**ssociation **P**re-test assessment, **E**chocardiographic and Natriuretic Peptide Score, **F**unctional Testing in Case of Uncertainty, and **F**inal Etiology) were calculated retrospectively using data abstracted from medical chart review, according to the reference literature (including age-specific cut-offs). Using the H2FPEF score, patients were categorized as having low (0 or 1), indeterminate (i.e., "intermediate"; 2–5), or high probability (6–9) of HFpEF. Using the HFA-PEFF score, patients were categorized as unlikely (0–1), indeterminate (i.e., "intermediate"; 2–4), or probable (5–6) HFpEF. The continuous probability of HFpEF was calculated for the H2FPEF score, but was not available for the HFA-PEFF score.

Statistical analysis

Baseline characteristics of included individuals at the time of index TTE were summarized using mean ± standard deviation for continuous variables (unless otherwise stated) and counts (proportions) for categorical variables, and compared between cases and controls using Students t tests and Fisher's exact tests respectively. The desired sample size was estimated according to modeling of estimated AI HFpEF model performance compared with previously identified clinical benchmarks, similar to previously published[17], and detailed further in the **Supplementary Methods**. Model diagnostic performance incorporated discrimination, calibration, classification, and clinical utility. Discrimination was assessed according to AUROC for models with continuous outputs (AI HFpEF model and H2FPEF), and compared via DeLong test[39]. Calibration of the models with continuous outputs was assessed visually according to calibration plots as well as evaluation of the intercept and slope of the flexible calibration curve and Hosmer-Lemeshow goodness-of-fit tests. Classification performance, including the sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) was determined for all models (Supplementary Table 12). Proportions tests are used for statistical comparisons between models for classification performance. Predictive values are reported according to the prevalence in the case-control design, and thus are expected to be higher than might be encountered in clinical practice; NPV and PPV are therefore reported at alternative prevalence(s), according to resampling of the current data, in Supplementary Fig. 2. Reclassification with application of the AI model, compared to existing clinical models, was assessed visually via alluvial plots and compared via a likelihood ratio test, and net reclassification statistics. Clinical utility was assessed using decision curve analysis[18], modeling the standardized net benefit and net reduction in interventions when using the clinical models, or clinical models in combination with the AI HFpEF model to make patient management decisions for a patient suspected of having HFpEF. Key interest was placed on the comparison of using only the existing clinical models, to a joint approach (positive by either the AI or clinical model), or conditional approach (positive by one model, or intermediate classification by one model and positive by another). Finally, the combination of all three models (AI HFpEF, H2FPEF, and HFA-PEFF via logistic regression), utilizing the most amount of information from the model (continuous output of AI HFpEF and H2FPEFF Score, categorical output of HFA-PEFF Score) was compared to assess incremental utility.

Prognostic performance was assessed via the association of model output (categorical and continuous) and patient outcomes. Kaplan-Meier curves were used to display time to death, HF hospitalization, and their composite by predicted classification for both AI HFpEF, H2FPEF, and HFA-PEFF scores and log-rank statistics computed. Cox proportional hazards models were used to estimate the unadjusted hazard ratio (HR) and 95% confidence interval (CI) for time to the outcome, separately according to predicted classes and quartiles of predicted risk according to continuous probabilities. Individuals not experiencing an outcome were censored at the last known study date. For the evaluation of HF hospitalization, Fine-Gray methods were used to account for the competing risk of death[40]. All analyses were performed using R version 4.2.3 (R Foundation, Vienna, Austria) using a two-tailed p value of <0.05 for significance unless otherwise stated. Bootstrap derived 95% confidence intervals are provided where appropriate.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the findings from this study are available within the manuscript and its supplementary information. Any additional raw data are available from the corresponding author upon reasonable request.

References

1. Martin, S. S. et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **149**, e347–e913 (2024).
2. Bozkurt, B. et al. HF STATS 2024: Heart Failure Epidemiology and Outcomes Statistics An Updated 2024 Report from the Heart Failure Society of America. *Journal of cardiac failure* https://doi.org/10.1016/j.cardfail.2024.07.001 (2024).
3. Oktay, A. A., Rich, J. D. & Shah, S. J. The emerging epidemic of heart failure with preserved ejection fraction. *Curr. Heart Fail Rep.* **10**, 401 (2013).

4. Reeves, G. R. et al. Comparison of frequency of frailty and severely impaired physical function in patients ≥60 years hospitalized with acute decompensated heart failure versus chronic stable heart failure with reduced and preserved left ventricular ejection fraction. *Am. J. Cardiol.* **117**, 1953–1958 (2016).
5. Warraich, H. J. et al. Physical function, frailty, cognition, depression, and quality of life in hospitalized adults ≥60 years with acute decompensated heart failure with preserved versus reduced ejection fraction. *Circ. Heart Fail* **11**, e005254 (2018).
6. Molloy, G. J., Johnston, D. W. & Witham, M. D. Family caregiving and congestive heart failure. Review and analysis. *Eur. J. Heart Fail* **7**, 592–603 (2005).
7. McDonagh, T. A. et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur. heart J.* **42**, 3599–3726 (2021).
8. Shah, S. J. et al. Research priorities for heart failure with preserved ejection fraction: National heart, lung, and blood institute working group summary. *Circulation* **141**, 1001–1026 (2020).
9. Obokata, M., Reddy, Y. N. V. & Borlaug, B. A. The role of echocardiography in heart failure with preserved ejection fraction: What do we want from imaging? *Heart Fail Clin.* **15**, 241–256 (2019).
10. Deaton, C., Edwards, D., Malyon, A. & MJ, S. Z. The tip of the iceberg: finding patients with heart failure with preserved ejection fraction in primary care. An observational study. *BJGP Open* **2**, bjgpopen18X101606 (2018).
11. Nagueh, S. F. et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J. Am. Soc. Echocardiogr.* **29**, 277–314 (2016).
12. Pieske, B. et al. How to diagnose heart failure with preserved ejection fraction: The HFA-PEFF diagnostic algorithm: a consensus recommendation from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC). *Eur. Heart J.* **40**, 3297–3317 (2019).
13. Reddy, Y. N. V., Carter, R. E., Obokata, M., Redfield, M. M. & Borlaug, B. A. A simple, evidence-based approach to help guide diagnosis of heart failure with preserved ejection fraction. *Circulation* **138**, 861–870 (2018).
14. Nikorowitsch, J. et al. Applying the ESC 2016, the H2FPEF, and the HFA-PEFF diagnostic algorithms for heart failure with preserved ejection fraction to the general population – a comparative approach. *Eur. Heart J.* **42**, ehab724.0856 (2021).
15. van de Bovenkamp, A. A. et al. Validation of the 2016 ASE/EACVI guideline for diastolic dysfunction in patients with unexplained dyspnea and a preserved left ventricular ejection fraction. *J. Am. Heart Assoc.* **10**, e021165 (2021).
16. Gevaert, A. B., Van De Heyning, C. M. & Tromp, J. Artificial intelligence to aid early detection of heart failure with preserved ejection fraction. *JACC: Adv.* **2**, 100447 (2023).
17. Akerman, A. P. et al. Automated echocardiographic detection of heart failure with preserved ejection fraction using artificial intelligence. *JACC: Adv.* **2**, 100452 (2023).
18. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **3**, 18 (2019).
19. James, S. L. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858 (2018).
20. Andersen, O. S. et al. Estimating left ventricular filling pressure by echocardiography. *J. Am. Coll. Cardiol.* **69**, 1937–1948 (2017).
21. Nauta, J. F. et al. Correlation with invasive left ventricular filling pressures and prognostic relevance of the echocardiographic diastolic parameters used in the 2016 ESC heart failure guidelines and in the 2016 ASE/EACVI recommendations: A systematic review in patients with heart failure with preserved ejection fraction. *Eur. J. Heart Fail* **20**, 1303–1311 (2018).
22. Nishimura, R. A. et al. Noninvasive doppler echocardiographic evaluation of left ventricular filling pressures in patients with cardiomyopathies: a simultaneous Doppler echocardiographic and cardiac catheterization study. *J. Am. Coll. Cardiol.* **28**, 1226–1233 (1996).
23. Ommen, S. R. et al. Clinical utility of Doppler echocardiography and tissue Doppler imaging in the estimation of left ventricular filling pressures: A comparative simultaneous Doppler-catheterization study. *Circulation* **102**, 1788–1794 (2000).
24. Jones, R. et al. Meta-analysis of echocardiographic quantification of left ventricular filling pressure. *ESC Heart Fail.* **8**, 566–576 (2021).
25. Rivas-Gotz, C., Manolios, M., Thohan, V. & Nagueh, S. F. Impact of left ventricular ejection fraction on estimation of left ventricular filling pressures using tissue Doppler and flow propagation velocity. *Am. J. Cardiol.* **91**, 780–784 (2003).
26. Chiou, Y. A., Hung, C. L. & Lin, S. F. AI-assisted echocardiographic prescreening of heart failure with preserved ejection fraction on the basis of intrabeat dynamics. *JACC Cardiovasc Imaging* **14**, 2091–2104 (2021).
27. Steyerberg, E. W. et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiol. (Camb., Mass.)* **21**, 128–138 (2010).
28. Administration, U. S. F. A. D. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*, https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (2019).
29. Cassianni, C. et al. Automated echocardiographic detection of heart failure with preserved ejection fraction using artificial intelligence is associated with cardiac mortality and heart failure hospitalization. *J. Am. Soc. Echocardiogr.* **37**, 914–916 (2024).
30. Khurshid, S. et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation* **145**, 122–133 (2022).
31. Hughes, J. W. et al. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *NPJ Digit Med* **6**, 169 (2023).
32. Borlaug, B. A., Sharma, K., Shah, S. J. & Ho, J. E. Heart failure with preserved ejection fraction: JACC scientific statement. *J. Am. Coll. Cardiol.* **81**, 1810–1834 (2023).
33. Heidenreich, P. A. et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **79**, e263–e421 (2022).
34. Fraiche, A. M. et al. Identification of Need for Ultrasound Enhancing Agent Study (the IN-USE Study). *Journal of the American Society of Echocardiography: official publication of the American Society of Echocardiography* https://doi.org/10.1016/j.echo.2020.07.015 (2020).
35. Lang, R. M. et al. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J. Am. Soc. Echocardiogr.* **28**, 1–39.e14 (2015).
36. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Bmj* **385**, e078378 (2024).
37. *Castor EDC*. https://castoredc.com (2019).
38. Harris, P. A. et al. The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inf.* **95**, 103208 (2019).

39. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).
40. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
41. Solomon, S. D. et al. Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *N. Engl. J. Med.* **387**, 1089–1098 (2022).

Acknowledgements

We would like to thank the participants and their families for their participation in the study. This study was supported by an investigator-initiated award from Ultromics Ltd. J.B.S. is supported by the National Institutes of Health (1R01HL169517 [J.B.S.], 1R01HL173998 [J.B.S.], 1R01AG063937 [J.B.S.]) and research grants (to the institution) from Anumana, Philips Healthcare, EVERSANA Lifesciences, and Bracco Diagnostics.

Author contributions

A.P.A., R.U, P.A.P., and J.B.S. conceived the project. N.A., C.A-J. and M.A.C. coordinated the project and assisted with acquisition of study data. N.A., C.A-J., M.A.C., R.T., L.B., and K.R. extracted TTE images and abstracted chart information. W.H., H.P., P.L. G.W., and R.U. provided funding and general study oversight. A.P.A. analyzed the images and clinical information. A.P.A. and J.B.S. authored the manuscript, with inputs from all other authors. All authors provided critical revisions to the manuscript.

Competing interests

J.B.S. reports investigator-initiated funding from Ultromics for the current study. J.B.S. reports research grants (to the institution) from Anumana, Philips Healthcare, EVERSANA Lifesciences, and Bracco Diagnostics; consulting for Bracco Diagnostics, Edwards Lifesciences, Philips Healthcare, General Electric Healthcare, and EVERSANA Lifesciences, and is a member of the scientific advisory boards for Ultromics, HeartSciences, Bristol Myers Squibb, Alnyam, and EchoIQ, and the data safety monitoring board for Pfizer. A.P.A., W.H., H.P., P.L., R.U., and G.W. are employees of Ultromics. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58283-7.

Correspondence and requests for materials should be addressed to Jordan B. Strom.

Peer review information *Nature Communications* thanks Shaan Khurshid, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.