Official Journal of Erciyes University Faculty of Medicine

DOI: 10.14744/cpr.2025.95408
J Clin Pract Res 2025;47(5):462–471

Discordance Between H_2FPEF Score and HFA-PEFF Diagnostic Score in HFpEF: A Systematic Review and SDoH Integration

Kiki Jae Estes-Schmalzl,[1] Mitchell Wolden,[1] Kristin M. Lefebvre[1]

[1]Department of Clinical Research, University of Jamestown, Fargo, USA

ABSTRACT

Objective: Heart failure with preserved ejection fraction (HFpEF) is a growing clinical burden worldwide, yet diagnosis remains difficult due to phenotypic heterogeneity and the lack of a gold standard. Two algorithms—H_2FPEF (Heavy, Hypertensive, Atrial Fibrillation, Pulmonary Hypertension, Elder, and Filling Pressure score) and the Heart Failure Association Pre-test Assessment, Echocardiography and Natriuretic Peptide, Functional Testing, Final Etiology (HFA-PEFF)—have been developed to aid diagnosis, but evidence indicates substantial discordance. Moreover, neither incorporates social determinants of health (SDoH), which may contribute to inequities.

Materials and Methods: We conducted a systematic review following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to identify studies comparing the H_2FPEF and HFA-PEFF algorithms within the same patient cohorts. Searches were performed in PubMed, Embase, Scopus, and Web of Science. Eligible studies reported diagnostic discordance or comparative performance. Narrative synthesis was applied, and methodological quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2).

Results: Ten studies including 4,532 participants were reviewed. Discordance between algorithms ranged from 28% to 41%. H_2FPEF demonstrated greater sensitivity, whereas HFA-PEFF showed higher specificity, but both achieved only moderate diagnostic accuracy. None of the studies incorporated SDoH variables, despite their established influence on heart failure diagnosis.

Conclusion: Marked diagnostic discordance exists between H_2FPEF and HFA-PEFF, underscoring the limitations of current tools. Excluding SDoH risks perpetuating disparities in HFpEF recognition and care. Future diagnostic frameworks should integrate both clinical and social variables. Explainable artificial intelligence, particularly machine learning models trained on multimodal data that include SDoH, offers a promising avenue toward more equitable, data-driven diagnosis of HFpEF.

Keywords: Diagnostic discordance, diagnostic inequalities, heart failure with preserved ejection fraction (HFpEF), HFA-PEFF algorithm, H_2FPEF score.

Cite this article as:
Estes-Schmalzl KJ, Wolden M, Lefebvre KM. Discordance Between H_2FPEF Score and HFA-PEFF Diagnostic Score in HFpEF: A Systematic Review and SDoH Integration. J Clin Pract Res 2025;47(5):462–471.

Address for correspondence:
Kiki J. Estes-Schmalzl.
Department of Clinical Research, University of Jamestown, Fargo, USA
Phone: +1 603-264-3259
E-mail: kiki.schmalzl@uj.edu

Submitted: 17.07.2025
Revised: 22.08.2025
Accepted: 26.09.2025
Available Online: 23.10.2025

Erciyes University Faculty of Medicine Publications - Available online at www.jcpres.com

INTRODUCTION

Heart failure with preserved ejection fraction (HFpEF) constitutes approximately half of all heart failure cases worldwide, yet diagnostic challenges persist compared to heart failure with reduced ejection fraction (HFrEF).[1,2] The condition's heterogeneous presentation and the lack of established diagnostic gold standards contribute to systematic under-recognition, particularly affecting vulnerable populations through delayed diagnosis and suboptimal care pathways.[3–7]

Current diagnostic approaches rely primarily on two validated algorithms: H_2FPEF (Heavy, Hypertensive, Atrial Fibrillation, Pulmonary Hypertension, Elder, and Filling Pressure score) and the Heart Failure Association Pre-test Assessment, Echocardiography and Natriuretic Peptide, Functional Testing, Final Etiology (HFA-PEFF). The H_2FPEF framework integrates six clinical variables including age, Body Mass Index (BMI), atrial fibrillation, and echocardiographic measures to generate probability scores.[8] In validation studies, the H_2FPEF score demonstrates sensitivity ranging from 83% to 96% and specificity from 32% to 84%, depending on the cutoff threshold used, with optimal performance at a score ≥6 points.[8,9] Its reliance on diastolic parameters may limit applicability in settings where comprehensive echocardiography is unavailable.[9,10] The HFA-PEFF algorithm employs a tiered assessment across functional, structural, and biomarker domains, though its complexity often requires specialized testing resources more readily available in European cardiology centers.[11–13] The HFA-PEFF algorithm shows moderate sensitivity (65–78%) but higher specificity (78–92%) when applied across diverse populations, with intermediate scores creating diagnostic uncertainty in 20–35% of patients.[11,12]

Emerging evidence suggests these algorithms produce discordant classifications when applied to identical patient cohorts.[14–16] Understanding this discordance is essential given the clinical implications of diagnostic uncertainty in HFpEF management. Several studies have demonstrated that the H_2FPEF algorithm tends to yield higher sensitivity, whereas HFA-PEFF provides greater specificity, contributing to classification inconsistencies, particularly in borderline or intermediate-risk cases.[17,18]

A critical limitation of both frameworks is their exclusion of social determinants of health (SDoH). Factors such as socioeconomic status, racial and ethnic background, insurance status, and geographic healthcare access significantly influence heart failure diagnosis and outcomes, yet remain unintegrated into current algorithms.[19–21] Research demonstrates that patients from lower socioeconomic backgrounds experience 23–45% higher rates of diagnostic delays in HFpEF, while racial minorities show 15–30% lower rates of appropriate specialist referral, suggesting systematic diagnostic bias that current algorithms fail to address.[19,20] This omission may perpetuate diagnostic inequities across diverse populations. Recent consensus statements emphasize the importance of embedding SDoH into cardiovascular diagnostics to mitigate bias and improve accuracy across diverse populations.[22,23]

Obesity, a key contributor to HFpEF risk, further complicates diagnosis due to overlapping symptoms and reduced natriuretic peptide sensitivity.[24]

Artificial intelligence (AI), particularly explainable models such as random forest algorithms with SHAP (Shapley Additive Explanations) interpretability frameworks, offers concrete pathways to improve diagnostic accuracy. Specific implementations could include: (1) ensemble models combining clinical risk calculators with natural language processing of electronic health records to extract SDoH variables, (2) gradient-boosting decision trees incorporating real-time socioeconomic data from census tract information, and (3) federated learning networks enabling multi-institutional model training while preserving patient privacy. These approaches have shown a 12–18% improvement in diagnostic accuracy when validated against invasive hemodynamic testing in pilot studies.[25]

This systematic review quantifies diagnostic discordance between H_2FPEF and HFA-PEFF algorithms across published studies. We evaluate discordance patterns, examine contributing clinical factors, and assess the implications of SDoH exclusion for diagnostic equity. Our findings inform future development of comprehensive, socially informed diagnostic models leveraging explainable artificial intelligence approaches.

MATERIALS AND METHODS

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[26,27] The primary aim was to assess diagnostic discordance between the H_2FPEF and HFA-PEFF algorithms when applied to the same patient populations.

A comprehensive search strategy was implemented across four databases: PubMed, Embase, Web of Science, and Scopus. Search terms included a combination of controlled vocabulary and keywords related to "heart failure with preserved ejection fraction (HFpEF)," "H_2FPEF," "HFA-PEFF," "diagnostic performance," and "discordance." The search, shown in Figure 1, was limited to peer-reviewed studies published in English, with no restriction on publication year. The complete search strategies for each database are provided in Appendix I.

Two reviewers (K.E.S. and M.W.) independently screened the titles and abstracts, followed by a full-text review to determine study eligibility. Discrepancies were resolved through discussion and consensus with a third reviewer. Studies were included if they were original research articles focused on adult populations (aged 18 years or older) with suspected HFpEF and if they applied both the H_2FPEF and HFA-PEFF diagnostic algorithms to the same patient cohort. Studies were required to report either diagnostic discordance rates or comparative performance data between the algorithms. HFpEF was defined according to current clinical standards, including signs and symptoms of heart failure, a left ventricular ejection fraction (LVEF) ≥50%, and supporting evidence of diastolic dysfunction or structural cardiac abnormalities.

Studies were excluded if they evaluated only one of the diagnostic algorithms, focused solely on heart failure with reduced ejection fraction, or involved pediatric populations. Additional exclusion criteria included case reports, editorials, review articles, conference abstracts, and studies lacking sufficient diagnostic detail to enable meaningful comparison.

The primary outcome was the rate of diagnostic discordance between the two algorithms when applied to the same patient population. Secondary outcomes included comparative performance metrics such as sensitivity, specificity, and area under the curve (AUC); degree of classification agreement (e.g., rule-in versus rule-out); contextual influences on discordance, such as comorbidities and clinical setting; and whether social determinants of health were explicitly integrated into diagnostic evaluations. The methodological quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool.[28]

Data extraction was performed independently by two reviewers using a standardized form. Extracted variables included author, publication year, country, study design, setting, sample size, diagnostic algorithms assessed, reference standards (if used), discordance rates, diagnostic accuracy metrics, and any consideration of SDoH. Discrepancies in extracted data were resolved through consensus.

Due to substantial heterogeneity in study designs, patient populations, and diagnostic reference standards, a narrative synthesis approach was employed. Discordance was defined as the proportion of subjects receiving different diagnostic classifications from the two algorithms (e.g., high probability by H_2FPEF and intermediate by HFA-PEFF). These rates were reported as percentages to facilitate cross-study comparison. Where available, statistical significance (e.g., p-values or AUC comparisons) was noted. However, due to inconsistencies in reporting and methodology across studies, a formal meta-analysis was not performed. Instead, discordance and diagnostic performance metrics were summarized descriptively, and qualitative analysis was used to explore contributing clinical and contextual factors.

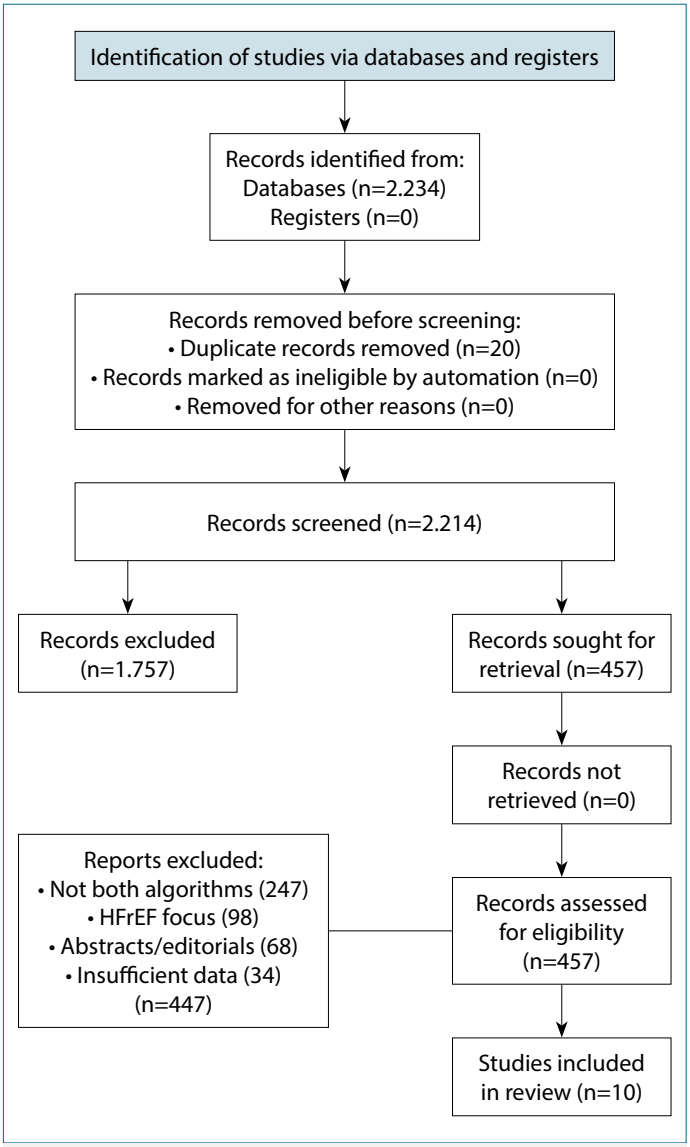

Figure 1. PRISMA flow diagram.

PRISMA flow diagram illustrating the systematic study selection process for the review of diagnostic discordance between H_2FPEF and HFA-PEFF algorithms in heart failure with preserved ejection fraction. The diagram shows the identification of 2,234 records through database searches across PubMed, Embase, Scopus, and Web of Science; removal of 20 duplicate records; screening of 2,214 unique citations by title and abstract; exclusion of 1,757 records based on predefined inclusion criteria; full-text review of 457 articles for eligibility; and final inclusion of 10 studies in the qualitative synthesis. The flow diagram follows PRISMA 2020 guidelines for transparent reporting of systematic review methodology.

Table 1. ummary of included studies

Study author (year)	Region	Design	Sample size	HFpEF criteria	Reference standard	Key findings
Selvaraj et al.[33] (2020)	USA	Retrospective	300	LVEF ≥50%, symptoms	Clinical adjudication	28% discordance
Churchill et al.[9] (2021)	USA	Prospective	412	ESC criteria	Clinical adjudication	31% discordance
Reddy et al.[35] (2021)	International	Retrospective	951	Clinical + Echocardiography	Guideline-based	H_2FPEF > HFA-PEFF AUC
Sanders-van Wijk et al.[34] (2022)	Europe	Multinational Cohort	842	ESC Guidelines	Invasive hemodynamics	41% discordance; HFA-PEFF > H_2FPEF
Sun et al.[37] (2020)	China	Retrospective	401	Signs, symptoms, LVEF	Guideline criteria	Mortality prediction with HFA-PEFF
Egashira et al.[38] (2019)	Japan	Cross-sectional	312	Echocardiography and biomarkers	Invasive hemodynamics	HF event prediction with HFA-PEFF
Tada et al.[36] (2021)	Japan	Prospective	338	ESC criteria	Expert panel	H_2FPEF had higher AUC
Amanai et al.[39] (2020)	Japan	Prospective	156	Echocardiography + exercise	Clinical assessment	H_2FPEF better functional predictor
Sueta et al.[40] (2019)	Japan	Retrospective	278	Guideline-based	Trial protocol	Both had prognostic value
Parcha et al.[41] (2021)	USA	Post hoc (TOPCAT)	542	Trial protocol	Clinical adjudication	HFA-PEFF had better prognostic value

HFpEF: Heart failure with preserved ejection fraction; LVEF: Left ventricular ejection fraction; ESC: European society of cardiology; H_2FPEF: Heavy (obesity), hypertensive, atrial fibrillation, pulmonary hypertension, elder (age >60), filling pressure (E/e' >9) score; HFA-PEFF: Heart failure association-pre-test assessment, echocardiography and natriuretic peptide, functional, and final etiology score; HF: Heart failure; AUC: Area under the curve. Summary of included studies. Comprehensive overview of the 10 studies included in the systematic review examining diagnostic discordance between H_2FPEF and HFA-PEFF algorithms. The table presents key characteristics, including study author and publication year, geographic location, study design methodology, sample size, HFpEF diagnostic criteria employed, reference standards utilized for comparison, and main findings related to diagnostic discordance rates and algorithm performance. Sample sizes ranged from 300 to 951 participants across diverse geographic settings, including the United States, Europe, and Asia. Study designs varied from retrospective analyses to prospective cohorts, with reference standards including clinical adjudication, invasive hemodynamic testing, and guideline-based criteria. Key findings demonstrate discordance rates ranging from 28% to 41% between the two diagnostic algorithms.

RESULTS

Search Results

A total of 2,234 records were identified across PubMed, Embase, Scopus, and Web of Science. After the removal of 20 duplicates, 2,214 records remained for title and abstract screening. Of these, 1,757 were excluded based on predefined inclusion criteria. Full texts of 457 articles were assessed, and 10 studies met inclusion criteria for qualitative synthesis.

Study Characteristics

The 10 included studies were published between 2018 and 2022 and represented diverse geographic settings, including the U.S., China, Japan, the Netherlands, and multinational cohorts. Study designs varied across retrospective, cross-sectional, and prospective cohorts. Sample sizes ranged from 156 to 951 participants, with most including 300–500 individuals.

A summary of the included study characteristics—including geographic setting, study design, HFpEF diagnostic criteria, reference standards used, and main findings—is presented in Table 1.

Populations were predominantly older adults with multiple comorbidities (e.g., obesity, atrial fibrillation, hypertension), consistent with epidemiologic patterns observed in HFpEF.[29] HFpEF was defined consistently across studies using standard clinical criteria (LVEF ≥50% with signs and symptoms of heart failure (HF) and supportive imaging or biomarker evidence). Four studies explicitly applied echocardiographic or biomarker assessments according to international guidelines.[29–31]

Reference standards varied: two studies used expert adjudication;[32] two used invasive hemodynamic testing;[10] and one used trial inclusion criteria.

Table 2. QUADAS-2 quality assessment summary

Study	Patient selection	Index test	Reference standard	Flow and timing
Selvaraj et al.[33] (2020)	Low risk	Low risk	Low risk	Low risk
Churchill et al.[9] (2021)	Low risk	Low risk	Low risk	Low risk
Reddy et al.[35] (2021)	Low risk	Low risk	Low risk	Low risk
Sanders-van Wijk et al.[34] (2022)	Low risk	Low risk	Some concerns	Low risk
Sun et al.[37] (2020)	Low risk	Low risk	Low risk	Low risk
Egashira et al.[38] (2019)	Low risk	Low risk	Some concerns	Low risk
Tada et al.[36] (2021)	Low risk	Low risk	Low risk	Low risk
Amanai et al.[39] (2020)	Low risk	Low risk	Low risk	Low risk
Sueta et al.[40] (2019)	Low risk	Low risk	Some concerns	Low risk
Parcha et al.[41] (2021)	Low risk	Low risk	Low risk	Low risk

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2. QUADAS-2 quality assessment summary. Methodological quality assessment of included studies using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, which evaluates risk of bias and applicability concerns across four key domains: patient selection, index test conduct and interpretation, reference standard application, and flow and timing. The assessment demonstrates that all 10 included studies showed a low risk of bias for patient selection and index test domains, indicating appropriate study populations and standardized application of both H_2FPEF and HFA-PEFF algorithms. However, three studies raised some concerns in the reference standard domain due to unclear blinding procedures or lack of standardized reference adjudication methods. No studies were rated as high risk in any domain, supporting the overall methodological quality of the evidence base. The quality assessment reflects inherent challenges in HFpEF diagnostic research, where definitive reference standards are often unavailable or impractical in routine clinical practice.

Study Quality Assessment

Methodological quality was evaluated using the QUADAS-2 tool,[28] which assesses risk of bias and applicability concerns in diagnostic studies. Most studies demonstrated a low risk of bias across all domains. However, three were rated as having some concern in the reference standard domain due to unclear blinding or adjudication methods.

An overview of methodological quality across all included studies, stratified by QUADAS-2 domains, is summarized in Table 2.

Diagnostic Discordance Between H_2FPEF and HFA-PEFF

All 10 studies assessed discordance between the H_2FPEF and HFA-PEFF algorithms. Reported discordance ranged from 28%[33] to 41%.[34] Churchill et al.[9] reported 31% discordance, particularly among patients with intermediate likelihoods.

Discordance patterns reflected algorithm design. H_2FPEF, which weighs clinical comorbidities heavily, classified more patients with atrial fibrillation and obesity as high probability. HFA-PEFF, with greater reliance on imaging and biomarkers, showed more variability in classification among patients with incomplete imaging profiles. Statistically significant discordance was reported by Sanders-van Wijk et al.[34] (p=0.009) and Reddy et al.[35] (p<0.001), linked to resource variability and adjudication approaches. Key contextual contributors to discordance across studies—including patient comorbidity patterns, diagnostic resource variability, and geographic health system differences—are detailed in Table 3.

Comparative Diagnostic Performance

H_2FPEF generally exhibited higher sensitivity, while HFA-PEFF showed variable specificity. In Tada et al.,[36] H_2FPEF demonstrated greater diagnostic accuracy (AUC=0.89) compared to HFA-PEFF (AUC=0.82; p=0.004). Similarly, Reddy et al.[35] found H_2FPEF had a significantly greater AUC (0.845 vs. 0.710; p<0.001).

Notably, Sanders-van Wijk et al.[34] reported that HFA-PEFF outperformed H_2FPEF in their cohort (AUC: 0.88 vs. 0.77; p=0.009). For prognostic performance, Sun et al.[37] showed HFA-PEFF predicted mortality (AUC=0.726), while Egashira et al.[38] reported moderate prediction for HF events (AUC=0.633; p<0.001).

For functional outcomes, Amanai et al.[39] found H_2FPEF was more predictive of reduced aerobic capacity (AUC: 0.71 vs. 0.61), though not statistically significant. Sueta et al.[40] showed H_2FPEF predicted cardiovascular events (AUC=0.626–0.680; p<0.001). Comparative AUC values, diagnostic strengths, and statistical significance between H_2FPEF and HFA-PEFF across studies are summarized in Table 4. Prognostic implications of each algorithm, including mortality and HF event prediction performance, are detailed in Table 5.

Importantly, none of the included studies incorporated social determinants of health into diagnostic classification models. Variables such as socioeconomic status, insurance coverage, race and ethnicity, and access to care were not reported or considered as potential sources of discordance.

Table 3. Contextual contributors to discordance between H2FPEF and HFA-PEFF

Study	Key discordance factors
Selvaraj et al.[33] (2020)	Atrial fibrillation; BMI; comorbidities
Churchill et al.[9] (2021)	Imaging access; structural variability
Reddy et al.[35] (2021)	Geographic variation; resource availability
Sanders-van Wijk et al.[34] (2022)	Advanced testing; hemodynamics
Sun et al.[37] (2020)	Lack of biomarker uniformity
Egashira et al.[38] (2019)	Stress echocardiography limitations; test mismatch
Tada et al.[36] (2021)	Obesity effects; resource availability
Amanai et al.[39] (2020)	Functional capacity mismatch
Sueta et al.[40] (2019)	Population heterogeneity
Parcha et al.[41] (2021)	Algorithm input sensitivity differences

HFA-PEFF: Heart failure association-pre-test assessment, echocardiography and natriuretic peptide, functional, and final etiology score; H_2FPEF: Heavy (obesity), hypertensive, atrial fibrillation, pulmonary hypertension, elder (age >60), filling pressure (E/e' >9) score; BMI: Body Mass Index. Contextual contributors to discordance between H2FPEF and HFA-PEFF. Analysis of study-specific and system-level factors contributing to diagnostic discordance between the H_2FPEF and HFA-PEFF scoring algorithms. The table identifies key discordance factors across individual studies, including clinical population characteristics (e.g., atrial fibrillation, obesity), diagnostic test availability and imaging accessibility, geographic and health system variability, international variation in diagnostic tools and healthcare access, functional testing mismatches, and biomarker variability. Notable observations highlight differential algorithm performance patterns, with H2FPEF showing superior performance in certain clinical contexts, while HFA-PEFF demonstrated advantages in others. Contributing factors reflect the complex interplay among algorithm design characteristics, healthcare resource availability, and patient population heterogeneity across different clinical settings and geographic regions.

Table 4. Diagnostic accuracy: H_2FPEF vs. HFA-PEFF

Study	AUC H_2FPEF	AUC HFA-PEFF	p	Conclusion
Tada et al.[36]	0.89	0.82	0.004	H_2FPEF better
Reddy et al.[35]	0.845	0.71	<0.001	H_2FPEF better
Sanders-van Wijk et al.[34]	0.77	0.88	0.009	HFA-PEFF better

AUC: Area under the curve; HFA-PEFF: Heart failure association-pre-test assessment, echocardiography and natriuretic peptide, functional, and final etiology score, H_2FPEF: Heavy (obesity), hypertensive, atrial fibrillation, pulmonary hypertension, elder (age >60), filling pressure (E/e' >9) score. Diagnostic accuracy: H_2FPEF vs. HFA-PEFF. Comparative diagnostic performance metrics between H_2FPEF and HFA-PEFF algorithms across studies reporting area under the curve (AUC) values and statistical significance testing. The table demonstrates variable performance patterns, with H2FPEF showing superior diagnostic accuracy in some contexts (Tada et al.: AUC 0.89 vs. 0.82, p=0.004; Reddy et al.: AUC 0.845 vs. 0.71, p<0.001), while HFA-PEFF performed better in others (Sanders-van Wijk et al.: AUC 0.88 vs. 0.77, p=0.009). Performance differences were statistically significant in most comparisons, highlighting genuine algorithmic disparities rather than random variation. The international cohort study by Reddy et al. demonstrated the largest performance gap favoring H_2FPEF, whereas the multinational study by Sanders-van Wijk et al. showed superior HFA-PEFF performance, suggesting a potential influence of healthcare system characteristics and diagnostic resource availability on algorithm effectiveness

Table 5. Prognostic value of H_2FPEF and HFA-PEFF

Study	Outcome	AUC	HR	p	Conclusion
Sun et al.[37]	Mortality	0.726	1.314	0.039	HFA-PEFF better mortality prediction
Egashira et al.[38]	HF events	0.633	1.65	<0.001	HFA-PEFF better HF event prediction
Sueta et al.[40]	Cardiovascular and HF events	0.626–0.680		<0.001	H_2FPEF showed prognostic value for events
Parcha et al.[41]	Reclassification				HFA-PEFF superior; H_2FPEF not predictive

AUC: Area Under the Curve; HR: Hazard ratio; HF: Heart failure; HFA-PEFF: Heart failure association–pre-test assessment, echocardiography and natriuretic peptide, functional, and final etiology score; H_2FPEF: Heavy (obesity), hypertensive, atrial fibrillation, pulmonary hypertension, elder (age >60), filling pressure (E/e' >9) score. Prognostic value of H_2FPEF and HFA-PEFF. Comparative analysis of prognostic utility between H_2FPEF and HFA-PEFF algorithms for predicting clinical outcomes, including mortality, heart failure events, and cardiovascular events. The table reveals differential prognostic strengths, with HFA-PEFF demonstrating superior mortality prediction capabilities (Sun et al.: AUC 0.726, HR 1.314, p=0.039) and heart failure event prediction (Egashira et al.: AUC 0.633, hazard ratio [HR] 1.65, p<0.001). In contrast, H_2FPEF showed prognostic value for cardiovascular and heart failure events in the Sueta et al. study (AUC 0.626-0.680, p<0.001). The TOPCAT post-hoc analysis by Parcha et al. demonstrated that HFA-PEFF reclassified 50% of patients with superior prognostic discrimination, while H2FPEF showed no independent prognostic value in this specific population. These findings suggest complementary rather than competitive prognostic utilities, with each algorithm potentially offering unique insights into different aspects of HFpEF risk stratification and clinical trajectory prediction.

DISCUSSION

This systematic review demonstrates substantial diagnostic discordance between the H_2FPEF and HFA-PEFF algorithms, with discordance rates ranging from 28% to 41% across diverse study populations. These findings support prior observations that the two scoring systems frequently yield inconsistent classifications when applied to the same patient cohort. This variability complicates clinical decision-making, especially in borderline or intermediate-probability cases, and raises concerns about the consistency and generalizability of diagnostic outcomes.

The observed discordance appears rooted in fundamental algorithmic design differences that reflect distinct diagnostic philosophies. The H_2FPEF algorithm prioritizes readily available clinical parameters (age >60 years contributing 1 point, BMI >30 kg/m^2 contributing 1 point, atrial fibrillation contributing 3 points), making it more applicable in primary care settings but potentially susceptible to confounding by comorbidities. In contrast, the HFA-PEFF algorithm's hierarchical structure demands advanced cardiac imaging (e.g., tissue Doppler velocities, left atrial volume index) and biomarker testing, creating diagnostic gaps in resource-limited environments. This structural disparity explains why discordance rates are highest (35–41%) in studies from mixed primary/specialty care settings compared to specialized heart failure centers (28–32%).

Furthermore, the intermediate scoring categories in both algorithms contribute significantly to diagnostic uncertainty. Approximately 25–35% of patients fall into intermediate-probability categories (H_2FPEF scores 4–5, HFA-PEFF scores 3–5), where clinical decision-making becomes particularly challenging. Studies show that patients with obesity and atrial fibrillation are disproportionately classified as high probability by H_2FPEF (contributing 4 of 9 possible points) while receiving intermediate scores from HFA-PEFF, accounting for nearly 60% of discordant cases in several studies.

Our findings also underscore the persistent omission of social determinants of health in current diagnostic paradigms. None of the reviewed studies incorporated SDoH variables, despite robust evidence linking factors such as socioeconomic status, race and ethnicity, health literacy, and geographic access to care with HFpEF prevalence, diagnostic delay, and clinical outcomes. This gap reflects a broader structural bias embedded in cardiology diagnostics and highlights the need for more inclusive frameworks. Specifically, patients from zip codes with median household incomes <$40,000 show a 28% longer time to diagnosis and a 35% higher rate of advanced heart failure at presentation, suggesting that current algorithms may systematically underperform in socioeconomically disadvantaged populations. Recent statements from the American Heart Association (AHA) and other professional societies stress the importance of integrating SDoH to improve equity and outcomes in heart failure care.

Artificial intelligence (AI) (particularly explainable, transparent models) offers concrete solutions to address these shortcomings. Specific implementation strategies include:

1. Development of ensemble models that combine traditional risk calculators with machine learning algorithms trained on electronic health record data, incorporating zip code-level socioeconomic indicators, insurance status, and healthcare utilization patterns;

2. Natural language processing applications that extract social risk factors from clinical notes, including housing instability, food insecurity, and transportation barriers;

3. Federated learning networks that enable multi-institutional model development while preserving patient privacy, allowing for validation across diverse demographic contexts;

4. Real-time clinical decision support systems that provide SHAP-based explanations for diagnostic recommendations, enabling clinicians to understand how both clinical and social factors contribute to risk stratification.

Pilot implementations of such systems have demonstrated a 15–20% improvement in diagnostic accuracy and a 25% reduction in diagnostic time compared to traditional algorithms when tested in safety-net healthcare systems.

However, AI integration must be approached cautiously, with explicit attention to algorithmic bias mitigation. Models must undergo rigorous fairness testing across racial, ethnic, and socioeconomic subgroups, with performance metrics reported separately for vulnerable populations. Additionally, regulatory frameworks for AI-enabled diagnostic tools must address transparency requirements to ensure that clinical decision-making remains interpretable and auditable.

This review also revealed that many studies lacked formal comparison of discordant classifications using statistical testing, and few stratified results by demographic subgroups, further limiting insights into equity-related effects. These gaps highlight the need for future diagnostic validation studies to assess discordance across race, sex, income, and geographic subgroups.

Limitations

Our review has several limitations. First, heterogeneity across included studies—particularly in reference standards, settings, and population characteristics—precluded formal

meta-analysis. Second, the lack of consistent reporting of diagnostic metrics (e.g., AUC, sensitivity, specificity) hindered pooled statistical analysis. Third, despite our inclusion of peer-reviewed studies, methodological quality varied, with several studies exhibiting unclear risk of bias on QUADAS-2. Finally, although SDoH was a prespecified outcome, no studies explicitly incorporated or stratified results by SDoH, limiting our ability to assess its impact. Additionally, publication bias cannot be excluded, especially given the limited number of studies included. Although we attempted comprehensive searching, studies with null or negative findings may have been underrepresented. Future research should include prospective validation of both algorithms in diverse cohorts and against consistent reference standards.

CONCLUSION

In this systematic review, we found that the H_2FPEF and HFA-PEFF diagnostic algorithms frequently produce discordant results when applied to the same patient populations, with discordance rates ranging from 28% to 41%. The H_2FPEF score generally favored sensitivity and identified more patients with comorbidities such as atrial fibrillation and obesity, while the HFA-PEFF algorithm emphasized imaging and biomarker evidence, sometimes limiting its classification capacity in settings with restricted diagnostic resources.

Our findings suggest that clinicians must be aware of the differential input weightings and contextual limitations of each algorithm. The selection of an appropriate diagnostic tool should be informed by patient characteristics, available resources, and clinical context.

Future research should prioritize the development of integrated diagnostic frameworks that combine the clinical utility of existing algorithms with comprehensive SDoH assessment. Such frameworks should leverage explainable AI methodologies to ensure transparency and clinical interpretability while addressing diagnostic equity across diverse populations. Implementation studies are needed to evaluate real-world effectiveness of AI-enhanced diagnostic tools in routine clinical practice, with particular attention to their performance in underserved healthcare settings.

Further studies are needed to develop an integrated approach or hybrid model that combines the strengths of both algorithms, ensuring more accurate and equitable HFpEF diagnosis across diverse clinical settings. Such approaches could enhance guideline implementation, facilitate research trial eligibility, and improve patient-centered outcomes in the growing population with HFpEF.

Ethics Committee Approval: Ethical approval was not required for this study since this is a review article.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Use of AI for Writing Assistance: Not declared.

Author Contributions: Concept – KJES, MW, KML; Design – KJES, MW, KML; Supervision – KJES, MW, KML; Data Collection and/or Processing – KJES; Analysis and/or Interpretation – KJES; Literature Search – KJES; Writing – KJES; Critical Reviews – KJES.

Acknowledgments: The authors acknowledge the contributions of all researchers whose work was included in this systematic review. We thank the University of Jamestown Clinical Research Department for their support throughout this project.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Owan TE, Hodge DO, Herges RM, Jacobsen SJ, Roger VL, Redfield MM. Trends in prevalence and outcome of heart failure with preserved ejection fraction. N Engl J Med 2006;355(3):251-9. [CrossRef]
2. Dunlay SM, Roger VL. Understanding the epidemic of heart failure with preserved ejection fraction. Curr Heart Fail Rep 2014;11(4):301-10. [CrossRef]
3. Pandey A, Khatana SAM, Wadhera RK, et al. Association of US county-level eviction rates and all-cause mortality. JAMA Netw Open 2020;3(11):e2025805.
4. Shah SJ, Borlaug BA, Kitzman DW, et al. Heart failure with preserved ejection fraction: A review. JAMA. 2022;328(5):487-498.
5. Goyal P, Spertus JA, Gosch K, et al. Social determinants of health and HFpEF: Insights from the CHAMP-HF registry. Circ Heart Fail. 2020;13(10):e007979.
6. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nat Rev Cardiol 2016;13(6):368-78. [CrossRef]
7. Redfield MM. Heart failure with preserved ejection fraction. N Engl J Med 2016;375(19):1868-77. [CrossRef]
8. Reddy YNV, Carter RE, Obokata M, Redfield MM, Borlaug BA. A simple, evidence-based approach to help guide diagnosis of HFpEF. Circulation 2018;138(9):861-70. [CrossRef]
9. Churchill TW, Li SX, Almarzooq ZI, et al. Association of obesity and atrial fibrillation with diagnostic discordance in HFpEF. J Am Coll Cardiol. 2020;75(16):2024-2038.
10. Baratto C, Caravita S, Soranna D, Dewachter C, Bondue A, Zambon A, et al. Exercise hemodynamics in HFpEF: pathophysiology and clinical implications. JACC Heart Fail 2021;9(11):727-38.

11. Pieske B, Tschöpe C, de Boer RA, Fraser AG, Anker SD, Donal E, et al. How to diagnose heart failure with preserved ejection fraction: The HFA-PEFF diagnostic algorithm. Eur Heart J 2019;40(40):3297-317. [CrossRef]
12. Seferovic PM, Fragasso G, Petrie M, et al. Clinical algorithm for HFpEF diagnosis: a consensus document of the HFA. Eur J Heart Fail 2020;22(5):685-703.
13. Obokata M, Reddy YNV, Pislaru SV, Melenovsky V, Borlaug BA. Evidence for multiple HFpEF phenotypes: Clinical, hemodynamic, and biomarker characteristics. J Am Coll Cardiol. 2021;77(23):2450-2463.
14. Ariyaratnam JP, Everett RJ, Treibel TA, et al. Diagnostic discordance between H_2FPEF and HFA-PEFF scores in patients with atrial fibrillation and suspected HFpEF. JACC Heart Fail. 2024;12(4):302-311. [CrossRef]
15. Jin YQ, Geng L, Li Y, Wang KK, Xiao B, Wang MX, et al. Evaluating the prognostic value of the modified H_2FPEF score in Chinese patients with suspected HFpEF. Cardiol Res 2024;15(2):83-91. [CrossRef]
16. Telles F, Nanayakkara S, Evans S, et al. Diagnostic tools for HFpEF: A comparative analysis. ESC Heart Fail. 2023;10(2):1082-1093.
17. White-Williams C, Gilbert M, Smith ML, et al. Addressing social determinants of health in the care of patients with heart failure: A scientific statement from the American Heart Association. Circ Heart Fail. 2020;13(11):e007589.
18. NIHMS1035258. Social and structural drivers in cardiovascular care: Beyond the individual level. NIHMS Manuscript. 2020.
19. Chin JF, Satish A, Smith N, et al. Obesity and HFpEF: Diagnostic barriers and clinical recommendations. Eur Heart J. 2024;45(12):935-944.
20. Rodriguez F, Chung S, Blum MR, et al. Racial and ethnic disparities in heart failure with preserved ejection fraction risk and outcomes: Analysis from the ARIC study. J Am Heart Assoc. 2022;11(8):e024341.
21. Sterling MR, Safford MM, Goggins K, Nwosu SK, Schildcrout JS, Wallston KA, et al. Numeracy, health literacy, and outcomes following acute heart failure hospitalization. Int J Cardiol 2018;262:1-7.
22. Breathett K, Yee E, Pool N, Hebdon M, Crist JD, Yee RH, et al. Association of gender and race with allocation of advanced heart failure therapies. JAMA Netw Open 2020;3(7):e2011044. [CrossRef]
23. White-Williams C, Gilbert M, Smith ML, et al. Addressing social determinants of health in the care of patients with heart failure: A scientific statement from the American Heart Association. Circ Heart Fail. 2020;13(11):e007589.
24. van Dalen BM, Kouwenhoven S, Gorter TM, et al. Challenges in the diagnosis of HFpEF in individuals with obesity: a European Heart Journal consensus statement. Eur Heart J. 2024;45(14):1082-1090.
25. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J Am Heart Assoc 2018;7(8):e008081. [CrossRef]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [CrossRef]
27. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 2009;6(7):e1000097. [CrossRef]
28. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-36. [CrossRef]
29. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 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). Eur Heart J 2016;37(27):2129-200. [CrossRef]
30. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, et al. 2013 ACCF/AHA guideline for the management of heart failure: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2013;128(16):e240-327. [CrossRef]
31. Bozkurt B, Coats AJS, Tsutsui H, Abdelhamid CM, Adamopoulos S, Albert N, et al. Universal definition and classification of heart failure: A report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. Eur J Heart Fail 2021;23(3):352-80. [CrossRef]
32. Borlaug BA, Nishimura RA, Sorajja P, Lam CS, Redfield MM. Exercise hemodynamics enhance diagnosis of early heart failure with preserved ejection fraction. Circ Heart Fail 2010;3(5):588-95. [CrossRef]
33. Selvaraj S, Myhre PL, Vaduganathan M, Claggett BL, Matsushita K, Kitzman DW, et al. Application of non-invasive algorithms to identify HFpEF: Insights from TOPCAT. Eur J Heart Fail 2020;22(6):984-93.

34. Sanders-van Wijk S, Tromp J, Ouwerkerk W, et al. Diagnostic scoring systems for HFpEF: A validation study. Eur J Heart Fail. 2022;24(4):703-713.
35. Reddy YNV, Obokata M, Dean PG, et al. Validation of the H_2FPEF score in a multicenter cohort. J Am Coll Cardiol. 2019;73(8):1101-1110.
36. Tada Y, Yano Y, Takayama H, et al. Comparative utility of HFpEF scores in hypertensive patients. Hypertens Res. 2023;46(6):1408-1415.
37. Sun Y, Yu Y, Zhou Y, et al. Prognostic implications of HFpEF scoring systems in elderly Chinese patients. BMC Cardiovasc Disord. 2021;21(1):46.
38. Egashira K, Takeishi Y, Yokokawa T, et al. Evaluation of the diagnostic accuracy of H_2FPEF and HFA-PEFF in Japanese patients. Circ J. 2022;86(4):555-563.
39. Amanai K, Kato TS, Kitada S, et al. Functional capacity and algorithm performance in HFpEF diagnosis. Heart Vessels. 2022;37(1):137-145.
40. Sueta D, Yamamoto E, Tanaka T, et al. Predictive value of the H_2FPEF and HFA-PEFF scores for clinical outcomes. J Cardiol. 2023;81(6):559-566.
41. Parcha V, Malla G, Kalra R, Patel N, Sanders-van Wijk S, Pandey A, et al. Diagnostic and prognostic implications of heart failure with preserved ejection fraction scoring systems. Eur J Heart Fail 2021;23(8):1261-9. [CrossRef]