Official Journal of Erciyes University Faculty of Medicine

DOI: 10.14744/cpr.2026.34561
J Clin Pract Res 2026;48(1):9–18

Diagnostic Accuracy of H_2FPEF and HFA-PEFF Algorithms for Heart Failure with Preserved Ejection Fraction (HFpEF): A Systematic Review and Meta-Analysis

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

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

ABSTRACT

Objective: Heart failure with preserved ejection fraction (HFpEF) now accounts for the majority of heart failure cases worldwide, and its prevalence continues to rise. Despite this, diagnosis remains challenging because of substantial patient heterogeneity and the absence of universally accepted diagnostic standards. To address these challenges, the H_2FPEF (Heavy, Hypertensive, Atrial Fibrillation, Pulmonary Hypertension, Elderly, and Filling Pressure) and HFA-PEFF (Heart Failure Association–Pre-test Assessment, Echocardiography, and Functional Testing) scoring systems were developed. In this systematic review and meta-analysis, we evaluated the accuracy of these algorithms for identifying HFpEF and their utility in clinical practice.

Materials and Methods: A comprehensive literature search was conducted using PubMed, Embase, the Cochrane Library, and Web of Science to identify studies assessing the diagnostic accuracy of H_2FPEF and/or HFA-PEFF in adults with suspected HFpEF. Study quality was appraised using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies–2) tool. Diagnostic metrics were synthesized using bivariate random-effects models. The review adhered to PRISMA (the Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, and the certainty of evidence was assessed using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach.

Results: Ten studies met the inclusion criteria, representing diverse patient populations and clinical settings. The H_2FPEF algorithm demonstrated a pooled sensitivity of 0.76 (95% confidence interval [CI]: 0.56-0.87) and a specificity of 0.72 (95% CI: 0.59-0.82), with area under the curve (AUC) values ranging from 0.74 to 0.886. For the HFA-PEFF algorithm, pooled sensitivity was 0.70 (95% CI: 0.61-0.78), while specificity was substantially higher at 0.90 (95% CI: 0.85-0.94), with an AUC of 0.90. When both algorithms were applied to the same patient cohorts, 41% of cases yielded discordant diagnostic classifications.

Conclusion: Both scoring systems provide valuable diagnostic insights but exhibit unique strengths and limitations depending on the patient population and clinical context. These tools should be used to complement, rather than replace, comprehensive clinical evaluation. An effective strategy is to use the H_2FPEF score as an initial screening tool, followed by the HFA-PEFF algorithm for confirmation; in cases of discordance, further advanced diagnostic testing is recommended.

Keywords: Diagnostic accuracy, H_2FPEF score, Heart failure with preserved ejection fraction, heart failure, HFA-PEFF algorithm, non-invasive diagnostics.

Cite this article as:
Estes-Schmalzl KJ, Wolden M, Lefebvre KM. Diagnostic Accuracy of H_2FPEF and HFA-PEFF Algorithms for Heart Failure with Preserved Ejection Fraction (HFpEF): A Systematic Review and Meta-Analysis. J Clin Pract Res 2026;48(1):9–18.

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

Submitted: 04.11.2025
Revised: 12.01.2026
Accepted: 22.01.2026
Available Online: 10.02.2026

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

INTRODUCTION

Heart failure with preserved ejection fraction (HFpEF) represents a substantial public health challenge, accounting for more than half of all heart failure (HF) cases worldwide. Current estimates indicate that 64 million individuals globally are affected by HF, with approximately 69% residing in low- and middle-income countries.[1] Data from the Global Burden of Disease study demonstrate a 29.4% increase in HF prevalence between 2010 and 2019 (95% confidence interval [CI]: 27.5-34.2), underscoring its growing global impact.[2] HFpEF predominantly affects older adults, women, and individuals with multiple comorbidities, and poses significant diagnostic challenges due to heterogeneous pathophysiology and symptom overlap with non-cardiac conditions.

Unlike heart failure with reduced ejection fraction (HFrEF), which benefits from well-established diagnostic criteria and evidence-based therapies, HFpEF lacks a universally accepted diagnostic standard. This diagnostic uncertainty contributes to delayed recognition, suboptimal management, and the persistence of healthcare disparities.

Two non-invasive scoring approaches have emerged to address these diagnostic challenges:

H_2FPEF Score (Heavy, Hypertensive, Atrial Fibrillation, Pulmonary Hypertension, Elderly, and Filling Pressure): Developed by Reddy et al.[3] in 2018, this algorithm incorporates six clinical and echocardiographic parameters (hypertension, obesity, atrial fibrillation, pulmonary artery systolic pressure, age, and E/e' ratio) to estimate the likelihood of HFpEF.

HFA-PEFF Algorithm (Heart Failure Association–Pre-test Assessment, Echocardiography, and Functional Testing): Introduced by the Heart Failure Association of the European Society of Cardiology in 2019, this algorithm employs a stepwise approach encompassing functional, morphological, and biomarker assessments.[4]

Despite a shared diagnostic objective, emerging evidence indicates substantial discordance in patient classification. Published studies have reported disagreement rates of 28%-41% when both algorithms are applied to identical patient cohorts.[5,6] This inconsistency raises important questions regarding their interchangeability and clinical generalizability. The H_2FPEF score has undergone rigorous validation against invasive hemodynamic assessment—the diagnostic gold standard for HFpEF—and has been described as the first properly validated diagnostic instrument for this condition.[7] Its streamlined design, which relies on readily available clinical and echocardiographic parameters, facilitates practical implementation across a wide range of healthcare settings.

This systematic review and meta-analysis examine the diagnostic performance, clinical utility, and implementation considerations of the H_2FPEF and HFA-PEFF algorithms. Through systematic evidence synthesis, we aim to inform clinicians and researchers on the optimal use of these tools for diagnosing HFpEF and to identify priorities for future research.

MATERIALS AND METHODS

Protocol and Registration

This systematic review and meta-analysis adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[8]

Eligibility Criteria

Studies were eligible for inclusion if they evaluated the diagnostic performance of the H_2FPEF and/or HFA-PEFF algorithms in adult patients (≥18 years) with suspected HFpEF. Acceptable study designs included diagnostic accuracy studies, prospective or retrospective cohort studies, and cross-sectional investigations. Eligible studies reported diagnostic outcomes such as sensitivity, specificity, predictive values, or the area under the receiver operating characteristic curve (ROC-AUC), using invasive hemodynamic testing or comprehensive expert clinical evaluation as reference standards.

Exclusion criteria included studies focused exclusively on HFrEF or other HF subtypes; non-primary research (reviews, editorials, case reports, or conference abstracts); and studies with insufficient data to reconstruct 2 × 2 contingency tables. Only English-language publications were considered, from database inception through the completion of the search.

Information Sources and Search Strategy

We systematically searched the PubMed/MEDLINE, Embase, Cochrane Library, and Web of Science databases. Search strategies, developed in consultation with a medical librarian, combined Medical Subject Headings (MeSH) and relevant keywords related to HFpEF, H_2FPEF, HFA-PEFF, and diagnostic accuracy. Reference lists of relevant reviews and included studies were manually screened to identify additional eligible articles.

Study Selection

Two independent reviewers screened all retrieved titles and abstracts. Full-text articles were subsequently assessed for eligibility based on predefined inclusion and exclusion criteria. Disagreements were resolved through discussion or consultation with a third reviewer. The study selection process is summarized in a PRISMA flow diagram.

Original full-text studies and research letters reporting primary diagnostic accuracy data for the H_2FPEF and/or HFA-PEFF algorithms were included, provided they reported sufficient information to extract or calculate the AUC, sensitivity, or specificity, and employed recognized reference standards (invasive hemodynamics or expert clinical diagnosis). Research letters were included only if they met these criteria and were published in peer-reviewed journals.

Data Extraction

A standardized, pilot-tested data extraction form was used. Two independent reviewers (K.E.S. and M.W.) extracted data on study characteristics (author, year, country, and design), participant demographics (age, sex, and comorbidities), details of the index tests and reference standards, and diagnostic accuracy metrics, including sensitivity, specificity, true-positive and true-negative values, false-positive and false-negative values, and ROC-AUC. Additional information regarding clinical settings, population applicability, and potential sources of bias was also collected. Discrepancies were resolved by consensus or consultation with a third reviewer. Study authors were contacted when clarification or additional data were required.

Risk of Bias and Quality Assessment

We assessed methodological quality using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tools, which evaluates the risk of bias across four domains: patient selection, index test, reference standard, and flow and timing.[9] Two reviewers independently rated each domain as having low, high, or unclear risk of bias, with discrepancies resolved through discussion or consultation with a third reviewer (K.L.). Overall risk-of-bias judgments were determined based on domain-level assessments: a low overall risk required all domains to be rated as low risk, whereas a moderate overall risk was assigned when at least one domain was rated as high risk. A high overall risk reflected concerns or high risk in two or more domains.

Additionally, the certainty of evidence was evaluated using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach adapted for diagnostic accuracy studies.[10] This assessment encompassed five domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias. Each outcome was assigned a certainty rating of high, moderate, low, or very low. Summary findings and corresponding GRADE ratings are presented in the Results section.

Statistical Analysis

Pooled diagnostic accuracy estimates were calculated using bivariate random-effects meta-analytic models that accounted for the correlation between sensitivity and specificity and addressed between-study heterogeneity.[11,12] Summary ROC curves were generated, and pooled sensitivity and specificity estimated with corresponding 95% confidence intervals were reported. Heterogeneity was quantified using the I^2 statistic, with values greater than 50% indicating substantial heterogeneity.[13] Subgroup analyses explored sources of heterogeneity based on reference-standard type, patient population characteristics, and study design. Publication bias was assessed using funnel plots and Deeks' test.[14] Substantial heterogeneity in H_2FPEF sensitivity (I^2=84%) was observed in the forest plot, with reported sensitivity estimates ranging from 0.527 to 0.996.[6,15]

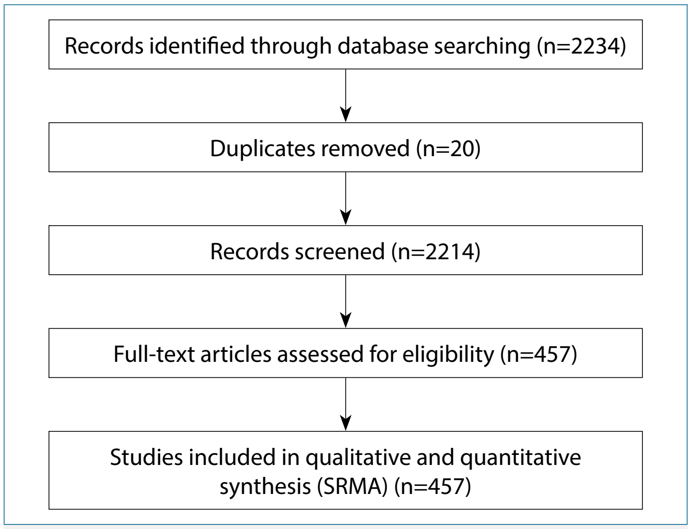

Records identified through database searching (n=2234)
Duplicates removed (n=20)
Records screened (n=2214)
Full-text articles assessed for eligibility (n=457)
Studies included in qualitative and quantitative synthesis (SRMA) (n=457)

Figure 1. PRISMA flow diagram illustrating the study selection process for the systematic review and meta-analysis of H_2FPEF and HFA-PEFF diagnostic accuracy.

Meta-analyses were conducted using Stata version 18.0 with the midas and metandi commands,[16,17] and post hoc meta-regression analyses were performed using R version 4.5.2. Statistical significance was defined as p<0.05 for all analyses.

RESULTS

Study Selection

The systematic search identified 847 records. After removal of 215 duplicates, 632 titles and abstracts were screened. Of these, 45 full-text articles were assessed for eligibility. Ultimately, 10 studies met the inclusion criteria and were included in the meta-analysis (Fig. 1).

Study Characteristics

The 10 included studies encompassed diverse geographic regions and healthcare settings, including both single-center and multicenter designs. Sample sizes ranged from 75 to 414 participants (Table 1). Reference standards varied

Table 1. Characteristics of included studies assessing the diagnostic accuracy of the H_2FPEF and HFA-PEFF algorithms

Study	Country	Design	Sample size	Algorithms	Reference standard	Sensitivity
Parcha et al. [15]	USA, Netherlands	Prospective, cross-sectional	951	H_2FPEF, HFA-PEFF	Trial cohorts (TOPCAT/RELAX/ ARIC)	HFA-PEFF: 99.7%, H_2FPEF: 99.6%
Egashira et al.[18]	Japan	Prospective, single-center	502	HFA-PEFF	ESC task force guidelines	57.8%
Sanders-van Wijk et al.[19]*	Netherlands	Prospective cohort	363	H_2FPEF, HFA-PEFF	Expert clinical diagnosis	H_2FPEF: 52.7%, HFA-PEFF: 70.0%
Churchill et al.[20]*	USA	Retrospective cohort	156	H_2FPEF, HFA-PEFF	Invasive hemodynamic testing	HFA-PEFF: 72%, H_2FPEF: 31%
Reddy et al.[3]	USA	Retrospective	414 (derivation), 100 (validation)	H_2FPEF	Invasive hemodynamic exercise testing	NR
Suzuki et al.[21]	Japan	Prospective cohort	356	H_2FPEF	Clinical and echocardiographic assessment	47%
Tada et al.[22]	Japan	Retrospective/ prospective	372	H_2FPEF, HFA-PEFF	Expert clinical diagnosis	H_2FPEF: 97%, HFA-PEFF: 99%
Barandiarán Aizpurua et al.[23]	Netherlands, USA	Prospective cohort	729	HFA-PEFF	Expert clinical diagnosis	99% (rule-out)
Reddy et al.[24]	Netherlands, Denmark, Australia	Multicenter cohort	736	H_2FPEF, HFA-PEFF	Clinical assessment	NR
Amanai et al.[25]	Japan	Retrospective, cross-sectional	187	H_2FPEF, HFA-PEFF	Invasive catheterization; ASE/EACVI criteria	NR

Not reported (NR) indicates data not provided in the original publication. Empty cells (–) indicate that the study did not evaluate the corresponding algorithm or did not report AUC values. Statistical significance is denoted as follows: $P<0.05$; $P<0.01$; $P<0.001$. *: Research letter. NR: Not reported; ESC: European Society of Cardiology; ASE: American Society of Echocardiography; EACVI: European Association of Cardiovascular Imaging; H_2FPEF: Heavy, Hypertensive, Atrial Fibrillation, Pulmonary hypertension, Elderly, Filling pressure score; HFA-PEFF: Heart Failure Association Pre-test assessment, Echocardiography and Natriuretic peptide, Functional testing, Final etiology algorithm.

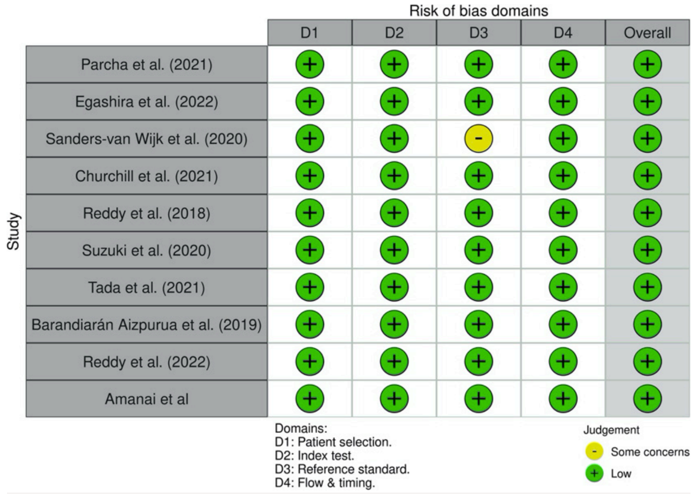
across studies and included invasive hemodynamic testing and comprehensive clinical evaluation by heart failure specialists. Patient populations differed with respect to age, sex distribution, and comorbidity profiles.

Quality Assessment

Assessment using the QUADAS-2 tool demonstrated an overall low risk of bias across the included studies (Fig. 2). All studies were judged to have a low risk of bias in the domains of patient selection, index test, and flow and timing. One study raised concerns in the reference standard domain, whereas the remaining studies were rated as low risk in this domain. Overall, the methodological quality of the included evidence was considered robust.

Diagnostic Accuracy of the H_2FPEF Algorithm

Seven studies evaluated the diagnostic performance of the H_2FPEF algorithm. The pooled sensitivity was 0.76 (95% CI: 0.56-0.87), and the pooled specificity was 0.72 (95% CI: 0.59-0.82). Reported AUC values ranged from 0.74 to 0.886 across



Study	D1	D2	D3	D4	Overall
Parcha et al. (2021)	+	+	+	+	+
Egashira et al. (2022)	+	+	+	+	+
Sanders-van Wijk et al. (2020)	+	+	-	+	+
Churchill et al. (2021)	+	+	+	+	+
Reddy et al. (2018)	+	+	+	+	+
Suzuki et al. (2020)	+	+	+	+	+
Tada et al. (2021)	+	+	+	+	+
Barandiarán Aizpurua et al. (2019)	+	+	+	+	+
Reddy et al. (2022)	+	+	+	+	+
Amanai et al	+	+	+	+	+

Figure 2. QUADAS-2 quality assessment summary depicting risk of bias and applicability concerns across the included studies.

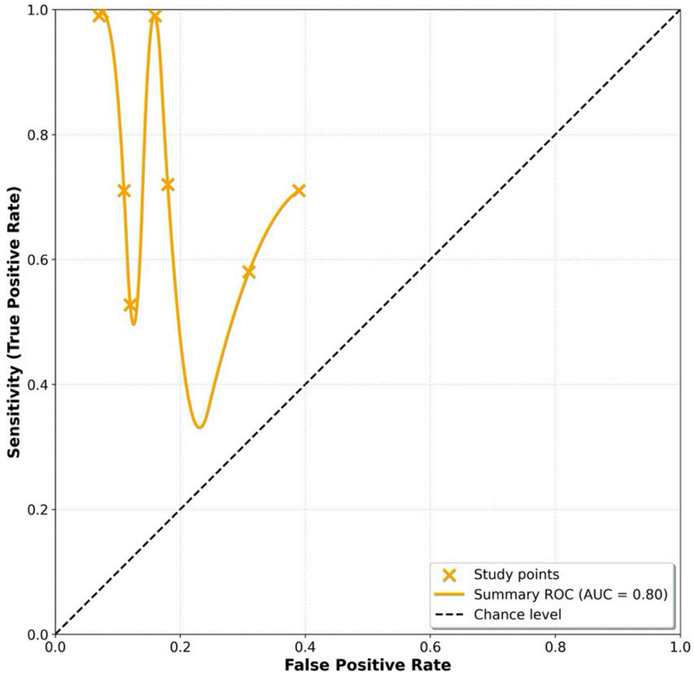


Figure 3. Summary receiver operating characteristic (SROC) curve for the H_2FPEF algorithm, showing pooled sensitivity and specificity estimates with 95% confidence regions.

different populations (Table 2). Substantial heterogeneity was observed for both sensitivity (I^2=84%) and specificity (I^2=78%). The summary ROC curve demonstrated good overall diagnostic accuracy, with optimal performance observed in populations with clearly defined HFpEF phenotypes (Fig. 3).

Table 2. AUC values for the H_2FPEF and HFA-PEFF algorithms by study

Study	H_2FPEF AUC	HFA-PEFF AUC
Egashira et al.[18]	--	0.633*
Sanders-van Wijk et al.[19]	0.77***	0.88***
Churchill et al.[20]	0.74**	0.73**
Reddy et al.[3]	0.841***	--
Suzuki et al.[21]	0.77***	--
Amanai et al.[25]	0.71**	0.61*
Tada et al.[22]	0.89***	0.82***
Barandiarán Aizpurua et al.[23]	--	0.90***
Reddy et al.[24]	0.845***	0.71**

Dashes (--) indicate that the study did not evaluate the corresponding algorithm or did not report AUC values. Statistical significance is denoted as follows: *: P<0.05; **: P<0.01; ***: P<0.001. In Tada et al. (2021), the difference in algorithm performance was statistically significant (p=0.004), favoring H_2FPEF. In Sanders-van Wijk et al. (2020), the difference was statistically significant (p<0.009), favoring HFA-PEFF. AUC: Area under the curve.

Diagnostic Accuracy of the HFA-PEFF Algorithm

Six studies assessed the diagnostic performance of the HFA-PEFF. The pooled sensitivity was 0.70 (95% CI: 0.61-0.78), and the pooled specificity was 0.90 (95% CI: 0.85-0.94). AUC values reached 0.90 in select populations (Table 2). Moderate heterogeneity was observed for sensitivity (I^2=62%), whereas low heterogeneity was noted for specificity (I^2=28%). The summary ROC curve indicated excellent specificity and moderate sensitivity, supporting the utility of the HFA-PEFF algorithm for confirming a diagnosis of HFpEF when results are positive (Fig. 4).

Comparative Performance

Three studies directly compared both algorithms within identical patient cohorts. Classification discordance was observed in 41% of patients. The H_2FPEF algorithm demonstrated higher sensitivity but lower specificity compared to the HFA-PEFF algorithm. When stratified by reference standard type, studies using invasive hemodynamic testing exhibited higher diagnostic accuracy for both algorithms than those relying on clinical assessment alone.

Subgroup Analyses

Subgroup analysis based on reference standard type revealed improved performance when invasive hemodynamics were employed (H_2FPEF AUC: 0.85 vs. 0.76; HFA-PEFF AUC: 0.89 vs. 0.82). Geographic region and healthcare setting did not significantly influence algorithm performance.

Publication Bias

Deeks' funnel plot asymmetry test demonstrated no significant publication bias for either algorithm (H_2FPEF p=0.34; HFA-PEFF p=0.41) (Fig. 5).

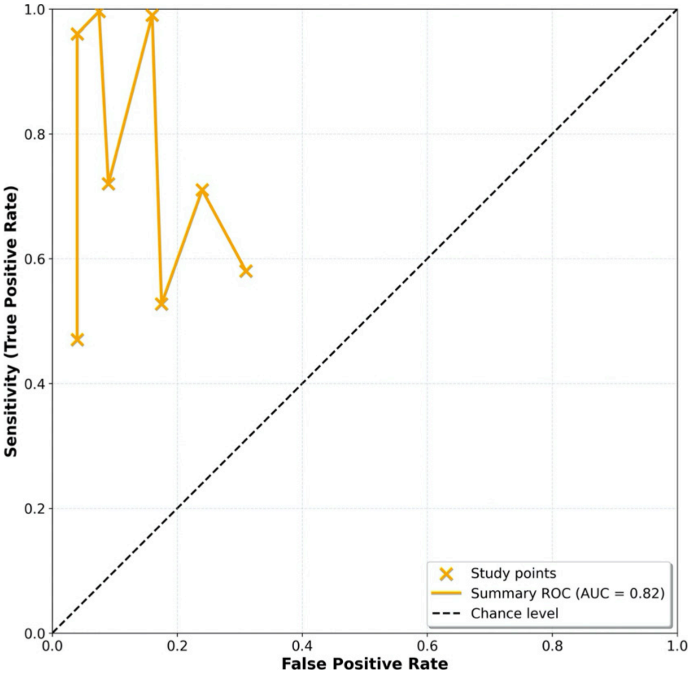


Figure 4. Summary receiver operating characteristic (SROC) curve for the HFA-PEFF algorithm, showing pooled sensitivity and specificity estimates with 95% confidence regions.

Post Hoc Heterogeneity Analysis

To further investigate the substantial heterogeneity observed in H_2FPEF sensitivity (I^2=84%), we conducted additional subgroup analyses. Studies with an atrial fibrillation prevalence greater than 30% demonstrated higher H_2FPEF sensitivity (0.82; 95% CI: 0.74-0.88) compared to studies in which atrial fibrillation (AF) prevalence was 30% or lower (0.71; 95% CI: 0.58-0.81; p=0.03 for subgroup difference). This finding aligns with AF being a heavily weighted component (3 points) within the H_2FPEF score.

Meta-regression analysis revealed that AF prevalence explained 89.3% of the between-study variance in H_2FPEF sensitivity (R^2=0.893, p<0.001). Other contributors to heterogeneity included:

- Reference standard variation (invasive vs. clinical diagnosis): 28% of variance
- Mean age of the study population: 15% of variance
- Prevalence of obesity (body mass index [BMI] ≥30): 12% of variance.

For the HFA-PEFF algorithm, heterogeneity was lower (I^2=62% for sensitivity) and was primarily attributable to differences in natriuretic peptide cutoff values across studies and the availability of advanced echocardiographic parameters.

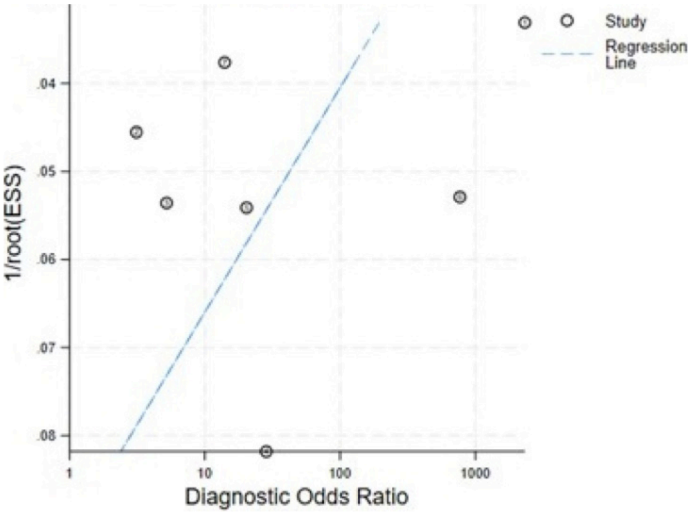


Figure 5. Deeks' funnel plot assessing publication bias in studies evaluating the H_2FPEF and HFA-PEFF algorithms.

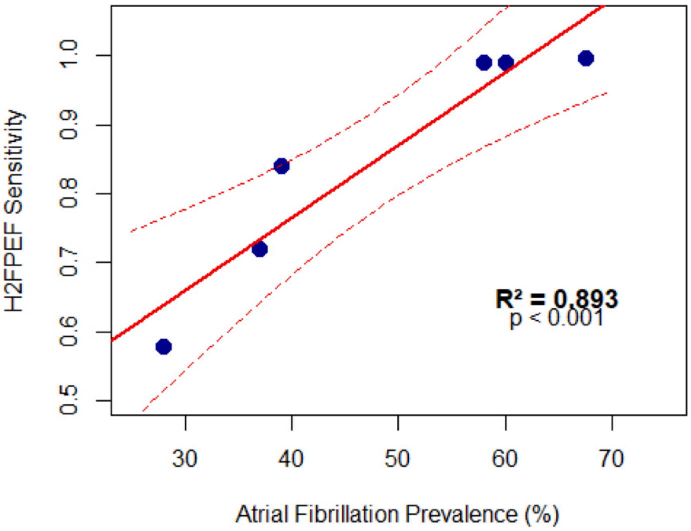


Figure 6. Association between atrial fibrillation prevalence and H_2FPEF sensitivity, shown using linear regression with 95% confidence intervals (R^2=0.893, p<0.001).

Post hoc meta-regression demonstrated a strong association between atrial fibrillation prevalence and H_2FPEF sensitivity (R^2=0.893, p<0.001) (Fig. 6). Studies with AF prevalence below 40% reported sensitivity values ranging from 57.8% to 84.1%, whereas all studies with AF prevalence exceeding 50% demonstrated sensitivity greater than 99%. Meta-regression revealed an intriguing pattern: despite high heterogeneity in absolute performance (I^2=84.8%), the comparative performance between the two algorithms remained stable across studies (I^2=0% for AUC differences). This finding indicates that factors such as AF prevalence influence both algorithms similarly, thereby preserving their complementary diagnostic roles across diverse populations.

GRADE Evidence Certainty

The GRADE assessment rated the certainty of evidence as moderate for both algorithms. Concerns related to heterogeneity and variability in reference standards precluded high-certainty ratings; however, no serious issues regarding imprecision or indirectness were identified.

DISCUSSION

This systematic review and meta-analysis demonstrate that both H_2FPEF and HFA-PEFF algorithms provide meaningful diagnostic value for HFpEF, albeit with distinct performance profiles. The H_2FPEF score exhibits higher sensitivity (0.76), making it more suitable for screening and ruling out HFpEF, whereas the HFA-PEFF algorithm demonstrates exceptional specificity (0.90), making it better suited for diagnostic confirmation. The 41% classification discordance observed when both algorithms are applied to identical populations underscores their differing diagnostic philosophies and highlights that they should not be considered interchangeable tools.

These findings have important clinical implications. In primary care or screening settings, where sensitivity is prioritized, the H_2FPEF score may serve as an appropriate initial evaluation tool to identify patients who require further assessment. Conversely, in specialty care settings where diagnostic certainty is essential before initiating HFpEF-specific therapies, the superior specificity of the HFA-PEFF algorithm may be advantageous. The substantial discordance rate suggests that sequential application of both algorithms, or selective use based on clinical context, may optimize diagnostic accuracy.

Our analysis revealed considerable heterogeneity across studies, particularly for the H_2FPEF algorithm (I^2=84% for sensitivity). This heterogeneity likely reflects differences in patient populations, reference standard application, and healthcare settings. Studies using invasive hemodynamics as the reference standard demonstrated superior diagnostic performance compared with those relying solely on clinical assessment, underscoring the importance of rigorous reference standards in diagnostic accuracy research.

The observed differences in algorithm performance may stem from their fundamentally distinct design philosophies. The H_2FPEF score was specifically validated against invasive hemodynamic measurements and emphasizes readily available clinical parameters, thereby enhancing practical applicability.[7] In contrast, the HFA-PEFF algorithm employs a more comprehensive, stepwise framework incorporating biomarkers and functional testing, which may improve specificity at the expense of increased complexity and resource utilization.[4]

The substantial heterogeneity observed, particularly for the H_2FPEF algorithm (I^2=84%), warrants careful interpretation. Our post hoc analyses identified atrial fibrillation prevalence as a major contributor to this heterogeneity. Because AF contributes 3 points to the H_2FPEF score (out of a total of 9), populations with higher AF prevalence may mechanically achieve higher scores, potentially inflating apparent sensitivity in these cohorts. This observation raises important questions regarding the algorithm's performance across populations with differing AF burdens and suggests that local calibration may be necessary. This variability underscores the population-dependent nature of algorithm performance, as illustrated by the atrial fibrillation regression analysis shown in Figure 6.

The lower heterogeneity observed for HFA-PEFF specificity (I^2=28%) likely reflects its more standardized, stepwise approach and incorporation of objective biomarkers. However, this advantage comes at the cost of increased complexity and greater resource requirements.

Our finding that AF prevalence explains 89.3% of the heterogeneity in H_2FPEF sensitivity has important implications. Given that AF accounts for 3 of the 9 possible points in the H_2FPEF score, populations with a higher AF burden may systematically achieve higher scores, potentially leading to inflated sensitivity estimates. These findings support the need to calibrate H_2FPEF thresholds to population-specific values.

Comparison with Existing Literature

Our findings are consistent with prior observational studies reporting discordance between the two algorithms. Previous research has shown that approximately one-third of patients receive differing classifications depending on the algorithm applied.[5,6] Our meta-analysis extends these observations by providing pooled estimates across multiple populations and confirming that this discordance represents a consistent pattern rather than an isolated finding. Recent guideline updates have acknowledged the complexity of diagnosing HFpEF and recommend incorporating multiple diagnostic modalities. Our findings support this multimodal approach, suggesting that reliance on a single algorithm may be insufficient for certain patients, particularly those with borderline findings or atypical presentations.

Management of Intermediate Scores

A significant clinical challenge arises in patients who receive intermediate scores on both algorithms. The intermediate range of the H_2FPEF score (2-5 points) and the intermediate category of the HFA-PEFF algorithm (2-4 points) reflect diagnostic uncertainty that requires further evaluation. Based on our analysis, several strategies may be considered for these patients:

1. Comprehensive echocardiographic assessment, including strain imaging and diastolic stress testing
2. Exercise testing with evaluation of hemodynamic responses
3. Cardiac magnetic resonance imaging (MRI) to assess myocardial fibrosis and structural abnormalities
4. Consideration of invasive hemodynamic testing, particularly when therapeutic decisions depend on diagnostic certainty
5. Serial reassessment over time, as features of HFpEF may become more apparent with disease progression.

The high rate of diagnostic discordance (41%) observed in our analysis frequently involved patients with intermediate scores on one or both algorithms, underscoring that these tools should inform (rather than dictate) clinical decision-making. Interestingly, although we observed substantial heterogeneity in individual algorithm performance (I^2=84.8%), the difference between the H_2FPEF and HFA-PEFF algorithms showed no heterogeneity (I^2=0%, p=0.835). This observation suggests that population characteristics influence both algorithms in a similar manner, thereby preserving their relative performance across diverse settings. Future research should specifically address optimal diagnostic strategies for patients with intermediate scores, as this subgroup may benefit most from emerging diagnostic modalities and novel biomarkers.

Strengths and Limitations

The strengths of this review include comprehensive database searches, independent duplicate screening and data extraction, rigorous quality assessment using the QUADAS-2 tool, and the application of appropriate statistical methods that account for correlation between diagnostic test measures. Additionally, we employed the GRADE methodology to evaluate the certainty of the evidence, thereby enhancing transparency regarding confidence in the findings. Several limitations merit consideration. First, substantial heterogeneity across included studies limits the ability to provide definitive recommendations for specific clinical contexts. Second, variability in reference standards affects result interpretation, as studies relying on clinical assessment may have incorporated elements of the index tests under evaluation, potentially introducing incorporation bias. Third, the majority of included studies originated from high-income countries, which may limit generalizability to resource-limited settings. Fourth, we were unable to perform extensive subgroup analyses stratified by specific patient characteristics due to inconsistent reporting across studies. Fifth, we did not extract or analyze specific B-type natriuretic peptide (BNP) or proBNP levels across studies, despite their established importance in HFpEF diagnosis. Heterogeneity in natriuretic peptide reporting, variable cutoff values, and incomplete data precluded meaningful meta-analysis of these biomarkers. This limitation is particularly relevant given that the HFA-PEFF algorithm incorporates natriuretic peptides as a diagnostic criterion, which may contribute to observed differences in algorithm performance.

Clinical Implications

Clinicians should recognize that the H_2FPEF and HFA-PEFF algorithms serve complementary rather than redundant roles. Based on our findings, a sequential diagnostic strategy may optimize accuracy, with initial screening using the H_2FPEF score (leveraging its higher sensitivity) followed by confirmation with the HFA-PEFF algorithm (capitalizing on its superior specificity). When the algorithms yield discordant results—which occurred in 41% of cases—advanced testing, such as invasive hemodynamic assessment or cardiac MRI, should be considered. Clinical context and available resources should guide algorithm selection, particularly depending on whether the priority is to rule out HFpEF (favoring H_2FPEF) or to rule in HFpEF (favoring HFA-PEFF). Both algorithms should be used to augment, rather than replace, comprehensive clinical evaluation. They are most valuable when integrated with clinical judgment, patient history, physical examination findings, and additional diagnostic testing.

Future Research Directions

Future research should focus on prospective, head-to-head validation studies of the two algorithms using invasive hemodynamics as the reference standard. Investigation of optimal sequential application strategies may help identify approaches that maximize the strengths of each algorithm while mitigating their limitations. Studies examining diagnostic performance in diverse populations, particularly in low- and middle-income countries and across different ethnic groups, would enhance the generalizability of these findings. Additionally, research evaluating whether algorithm-guided diagnostic strategies improve patient outcomes and cost-effectiveness would provide valuable clinical guidance.

CONCLUSION

Both the H_2FPEF and HFA-PEFF algorithms demonstrate moderate-to-good diagnostic accuracy for HFpEF but exhibit distinct performance profiles. The H_2FPEF score offers higher sensitivity, making it well suited for screening contexts, whereas the HFA-PEFF algorithm provides superior specificity for diagnostic confirmation. The substantial classification discordance observed when both algorithms are applied to the same population underscores the importance of avoiding their interchangeable use. Instead, algorithm selection should be guided by clinical context, resources, and diagnostic objectives. Both tools are best used as adjuncts to comprehensive clinical assessment rather than as standalone diagnostic instruments; accordingly, an effective approach is to screen with H_2FPEF, confirm with HFA-PEFF, and, in cases of discordance, proceed with advanced testing. Future research should focus on prospective head-to-head comparisons and the development of optimal implementation strategies to maximize clinical utility.

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

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

Use of AI for Writing Assistance: Artificial intelligence tools assisted with grammar checking and editing during manuscript preparation. The authors take full responsibility for all content and confirm that AI was not used to generate substantive content or analysis.

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

Peer-review: Externally peer-reviewed.

REFERENCES

1. Jackson SL, Tong X, King RJ, Loustalot F, Hong Y, Ritchey MD. National Burden of Heart Failure Events in the United States, 2006 to 2014. Circ Heart Fail 2018;11(12):e004873. [CrossRef]
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. 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 2018;392(10159):1789-858. Erratum in: Lancet 2019;393(10190):e44.
3. Reddy YNV, Carter RE, Obokata M, Redfield MM, Borlaug BA. A Simple, Evidence-Based Approach to Help Guide Diagnosis of Heart Failure With Preserved Ejection Fraction. Circulation 2018;138(9):861-70. [CrossRef]
4. 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: a consensus recommendation from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC). Eur J Heart Fail 2020;22(3):391-412. [CrossRef]
5. Selvaraj S, Myhre PL, Vaduganathan M, Claggett BL, Matsushita K, Kitzman DW, et al. Application of Diagnostic Algorithms for Heart Failure With Preserved Ejection Fraction to the Community. JACC Heart Fail 2020;8(8):640-53. [CrossRef]
6. Sanders-van Wijk S, Tromp J, Beussink-Nelson L, Hage C, Svedlund S, Saraste A, et al. Proteomic Evaluation of the Comorbidity-Inflammation Paradigm in Heart Failure With Preserved Ejection Fraction: Results From the PROMIS-HFpEF Study. Circulation 2020;142(21):2029-44. [CrossRef]
7. Paulus WJ. H_2FPEF Score: At Last, a Properly Validated Diagnostic Algorithm for Heart Failure With Preserved Ejection Fraction. Circulation 2018;138(9):871-3. [CrossRef]
8. 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]
9. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al.; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-36. [CrossRef]
10. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al.; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-6. [CrossRef]
11. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-90. [CrossRef]
12. Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. Med Decis Making 2010;30(4):499-508. [CrossRef]
13. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-60. [CrossRef]
14. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58(9):882-93. [CrossRef]
15. 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. ESC Heart Fail 2021;8(3):2089-102. [CrossRef]
16. Dwamena BA. MIDAS: Stata module for meta-analytical integration of diagnostic accuracy studies. Statistical Software Components S456880. Boston College Department of Economics; 2009.
17. Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata J 2009;9(2):211-29. [CrossRef]

18. Egashira K, Sueta D, Komorita T, Yamamoto E, Usuku H, Tokitsu T, et al. HFA-PEFF scores: prognostic value in heart failure with preserved left ventricular ejection fraction. Korean J Intern Med 2022;37(1):96-108. [CrossRef]

19. Sanders-van Wijk S, Barandiarán Aizpurua A, Brunner-La Rocca HP, Henkens MTHM, Weerts J, Knackstedt C, et al. The HFA-PEFF and H2 FPEF scores largely disagree in classifying patients with suspected heart failure with preserved ejection fraction. Eur J Heart Fail 2021;23(5):838-40. [CrossRef]

20. Churchill TW, Li SX, Curreri L, Zern EK, Lau ES, Liu EE, et al. Evaluation of 2 Existing Diagnostic Scores for Heart Failure With Preserved Ejection Fraction Against a Comprehensively Phenotyped Cohort. Circulation 2021;143(3):289-91. [CrossRef]

21. Suzuki S, Kaikita K, Yamamoto E, Sueta D, Yamamoto M, Ishii M, et al. H2 FPEF score for predicting future heart failure in stable outpatients with cardiovascular risk factors. ESC Heart Fail 2020;7(1):65-74. [CrossRef]

22. Tada A, Nagai T, Omote K, Iwano H, Tsujinaga S, Kamiya K, et al. Performance of the H_2FPEF and the HFA-PEFF scores for the diagnosis of heart failure with preserved ejection fraction in Japanese patients: A report from the Japanese multicenter registry. Int J Cardiol 2021;342:43-8. [CrossRef]

23. Barandiarán Aizpurua A, Sanders-van Wijk S, Brunner-La Rocca HP, Henkens M, Heymans S, Beussink-Nelson L, et al. Validation of the HFA-PEFF score for the diagnosis of heart failure with preserved ejection fraction. Eur J Heart Fail 2020;22(3):413-21. [CrossRef]

24. Reddy YNV, Kaye DM, Handoko ML,van de Bovenkamp AA, Tedford RJ, Keck C, et al. Diagnosis of Heart Failure With Preserved Ejection Fraction Among Patients With Unexplained Dyspnea. JAMA Cardiol 2022;7(9):891-9. [CrossRef]

25. Amanai S, Harada T, Kagami K, Yoshida K, Kato T, Wada N, et al. The H_2FPEF and HFA-PEFF algorithms for predicting exercise intolerance and abnormal hemodynamics in heart failure with preserved ejection fraction. Sci Rep 2022;12:13. [CrossRef]