Discordance Between H2FPEF and HFA-PEFF Diagnostic Scores in HFpEF: A Systematic Review and Call for SDoH Integration

Authors: [1]Kiki J. Estes-Schmalzl, BSN, MBA-MIS, RN, Ph.D. Candidate, University of Jamestown Clinical Research, (kiki.schmalzl@uj.edu OR kiki.estes@gmail.com), (Phone: 603-264-3259), https://orcid.org/0009-0002-0725-5113

[2]Mitchell Wolden, PhD, DPT, Associate Director of Clinical Research at the University of Jamestown. (mwolden@uj.edu). (Phone: 701-356-2136), https://orcid.org/0000-0002-9390-1590

[3]Kristin M. Lefebvre, PT, PhD, Director of Clinical Research at the University of Jamestown, (kristin.lefebvre@uj.edu), (Phone: 302-561-0787), 6000 College Lane. Jamestown, ND 58405

Corresponding Author: [1]Kiki J. Estes-Schmalzl, Email: kiki.schmalzl@uj.edu, Phone: 603-264-3259

Postal address: Department of Clinical Research, University of Jamestown, 4190 26th Ave. S. Fargo, ND 58104, United States of America

Abstract

Background: Heart failure with preserved ejection fraction (HFpEF) represents a growing burden worldwide, yet diagnosis remains challenging due to clinical heterogeneity and the absence of a gold standard. Two diagnostic algorithms, H2FPEF and HFA-PEFF, have been developed to assist in HFpEF diagnosis; however, emerging evidence suggests substantial discordance between these two approaches. The omission of social determinants of health (SDoH) from existing diagnostic frameworks may further exacerbate diagnosis difficulties.

Methods: A systematic review was conducted according to PRISMA guidelines. Searches were performed in PubMed, Embase, Web of Science, and Scopus databases using controlled vocabulary related to HFpEF diagnostic algorithms. Studies were included if they applied the H2FPEF and HFA-PEFF algorithms within the same patient cohort and reported comparative diagnostic performance or discordance rates. A narrative synthesis approach was used for qualitative analysis. Methodological quality was assessed using the QUADAS-2 tool.

Results: Ten studies met the inclusion criteria, encompassing 4,532 subjects across diverse populations and settings. Discordance rates between H2FPEF and HFA-PEFF scores ranged from 28% to 41%. The H2FPEF score demonstrated higher sensitivity, whereas the HFA-PEFF algorithm exhibited greater specificity. None of the studies incorporated SDoH variables into diagnostic evaluations. Variability in patient comorbidities, echocardiographic parameters, and clinical settings contributed to diagnostic discordance.

Conclusion: Substantial discordance between current HFpEF diagnostic algorithms highlights limitations in existing rule-based tools. The absence of SDoH integration into current frameworks may perpetuate diagnostic inequities. Future diagnostic models should incorporate clinical and social risk factors to enhance diagnostic accuracy and promote health equity, with explainable artificial intelligence (AI) offering a promising pathway forward.

Keywords

- Heart failure with preserved ejection fraction (HFpEF)
- H2FPEF score
- HFA-PEFF algorithm
- Diagnostic discordance
- Social determinants of health (SDoH)
- Explainable artificial intelligence (XAI)
- Diagnostic equity
- Heart failure diagnosis

Introduction

Heart failure with preserved ejection fraction (HFpEF) represents a growing clinical and public health challenge, accounting for over 50% of all heart failure (HF) cases worldwide [1, 2]. Despite its prevalence, HFpEF remains underdiagnosed and undertreated compared to heart failure with reduced ejection fraction (HFrEF), primarily due to its diagnostic complexity, clinical heterogeneity, and the absence of a universally accepted gold standard [3 - 5]. Accurate diagnosis of HFpEF is critical, as delayed or missed identification can exacerbate morbidity, hospital readmissions, and healthcare disparities, particularly among socially vulnerable populations [6, 7].

Two non-invasive diagnostic algorithms, the H2FPEF and the HFA-PEFF, have been developed to improve diagnostic accuracy [5, 8]. The H2FPEF algorithm leverages six routinely collected clinical and echocardiographic variables, such as age, body mass index, atrial fibrillation, and E/e' to estimate the HFpEF likelihood [5]. However, the completeness of echocardiographic parameters like E/e' can vary across clinical environments. Prior studies have reported incomplete capture of diastolic indices in non-specialty or community settings, which may limit the routine applicability of the H2FPEF score [9, 10].

In contrast, the HFA-PEFF algorithm, endorsed by the Heart Failure Association of the European Society of Cardiology (ESC), utilizes a structured, stepwise approach incorporating functional, morphological, and biomarker domains [8]. Compared to H2FPEF, it often requires advanced imaging and natriuretic peptide testing, making it more widely used in European heart failure centers [11, 12].

Although both algorithms have been validated in prospective research, their real-world application remains variable. Early comparative studies suggest meaningful classification differences may arise when the same cohort is assessed using both algorithms [13 - 15]. This systematic review aims to evaluate these discordance rates and the clinical and contextual factors—such as differences in score weighting, comorbidity profiles, and access to diagnostics—that shape discordant classifications.

Notably, both diagnostic algorithms omit social determinants of health (SDoH), including socioeconomic status, race/ethnicity, insurance coverage, and geographic access to care [16 - 18]. These factors substantially influence diagnostic access and treatment patterns in heart failure (HF) and may contribute to disparities in HFpEF diagnosis. Addressing this omission is critical for developing more equitable diagnostic strategies.

This review synthesizes current evidence on the diagnostic discordance between the H2FPEF and HFA-PEFF algorithms in subjects with suspected HFpEF. Our primary objective is to quantify discordance rates across diverse clinical populations. We also examine the role of SDoH omission and propose directions for future context-aware diagnostic models, including the application of explainable artificial intelligence (AI) to improve diagnostic equity.

Methods

This systematic review was conducted as a secondary analysis under the PROSPERO-registered protocol [19]. It focused on diagnostic discordance between HFpEF diagnostic algorithm performance when applied to the same patient populations. Our systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [20].

Data Sources and Search Strategy

Our systematic search was conducted in PubMed, Embase, Web of Science, and Scopus databases. The full search strategy for each database is included in Appendix I. The search combined controlled vocabulary terms and keywords related to HFpEF, H2FPEF, HFA-PEFF, diagnostic classification, and discordance. It was restricted to peer-reviewed articles published in English, with no date restrictions.

Two independent reviewers (K.E.S. and M.W.) screened all titles and abstracts, followed by a full-text review of potentially eligible studies. A third reviewer resolved discrepancies through discussion or adjudication.

Study Selection

Studies were included in this review if they satisfied specific inclusion criteria. Eligible studies were original research articles published in peer-reviewed journals and focused on adult populations (≥18 years) with suspected heart failure with preserved ejection fraction (HFpEF). A key requirement for inclusion was the application of both the H2FPEF and HFA-PEFF diagnostic algorithms within the same patient cohort. Additionally, studies needed to report either diagnostic discordance rates or provide comparative diagnostic data between the two algorithms.

HFpEF was defined in accordance with contemporary clinical standards. This included the presence of signs and symptoms of heart failure, a left ventricular ejection fraction (LVEF) of 50% or greater, and evidence of either structural heart disease or diastolic dysfunction.

Studies were excluded if they assessed only one of the diagnostic algorithms (either H2FPEF or HFA-PEFF) without any direct comparison. Further exclusion criteria encompassed studies focusing solely on heart failure with reduced ejection fraction (HFrEF), those involving pediatric populations, and publications such as case reports, editorials, conference abstracts, review articles, or any studies lacking sufficient diagnostic detail to allow meaningful comparison.

Outcomes

Outcomes categories were defined a priori in the registered review protocol. Table 1 summarizes the key outcomes assessed in the review and the rationale for their inclusion. The primary outcome was the rate of diagnostic discordance between the H2FPEF and HFA-PEFF algorithms when applied to the same patient population. Secondary outcomes included: (1) comparative diagnostic performance metrics (e.g., sensitivity, specificity, AUC); (2) agreement or disagreement in classification outcomes (e.g., rule-in vs. rule-out); (3) influence of population-level factors on discordance (e.g., comorbidities, setting); and (4) presence or absence of SDoH integration. Methodological quality was also evaluated using QUADAS-2 [22].

Table 1: Summary of Predefined Study Outcomes and Their Rationale

Outcome Category	Outcome Description	Rationale or Relevance
Primary Outcome	Diagnostic discordance rates between H2FPEF and HFA-PEFF	Addresses potential inconsistencies in current tools
Secondary Outcome	Comparative diagnostic performance (sensitivity, specificity, AUC)	Assesses trade-off between detection vs. precision
Secondary Outcome	Contributing factors to discordance (e.g., obesity, AF, imaging access)	Highlight's role of patient and system-level variability
Secondary Outcome	Study quality assessed via QUADAS-2	Informs strength of evidence across studies

Interpretive Outcome	Consideration of SDoH integration	Identifies gaps in diagnostic equity frameworks

Note: Outcome categories were defined a priori in the registered review protocol to ensure comprehensive evaluation of diagnostic algorithm performance and contextual factors influencing discordance.

Data Extraction

Data extraction was performed using a standardized form capturing the following variables: author, publication year, country, study design and setting, sample size, diagnostic algorithms evaluated, reference standard used, discordance rates, diagnostic accuracy metrics (sensitivity, specificity, AUC), and mention of SDoH considerations. Two reviewers independently extracted data, and discrepancies were reconciled through consensus.

Quality Assessment

Methodological quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool [22]. This tool was selected due to its widespread use and methodological rigor in diagnostic accuracy research. The QUADAS-2 tool assesses study quality across four domains: patient selection, index test, reference standard, and flow and timing. Each domain was rated for risk of bias as 'low' or some concerns based on review of study methods and reporting.

Importantly, QUADAS-2 does not generate an overall summary score; no such scores were calculated in this review. Instead, domain-level judgments were used to provide a more accurate and context-sensitive evaluation of methodological quality. Applicability concerns were also assessed for the first three domains to determine whether the included studies were generalizable to the review question.

Narrative Synthesis

Given the heterogeneity in study designs, patient populations, and reference standards, a narrative synthesis approach was used. Discordance between diagnostic algorithms was defined as the proportion of subjects classified differently (e.g., high vs. intermediate probability) by H2FPEF versus HFA-PEFF algorithms when applied within the same cohort. We calculated this discordance for each study by identifying cases where a subject received different diagnostic classifications from the two algorithms.

Discordance rates were reported as percentages to enable standardized comparison across studies with varying sample sizes. While several studies did report p-values for comparing diagnostic performance metrics of individual algorithms, few reported statistical significance testing for direct discordance between algorithms. Where available, these comparative statistics were incorporated into our synthesis. However, the variability in reporting formats and inconsistent statistical approaches across studies precluded formal meta-analysis of discordance rates. Instead, discordance rates and diagnostic performance metrics were summarized descriptively, while contextual factors contributing to discordance were examined qualitatively.

Results

Search Results

A total of 2,234 records were identified through database searches across PubMed, Embase, Scopus, and Web of Science. After the removal of 20 duplicate records, 2,214 unique citations were screened by title and abstract. Of these, 1,757 records were excluded for irrelevance based on predefined inclusion criteria. The full texts of 457 articles were reviewed for eligibility. After applying inclusion and exclusion criteria, 10 studies were included in the qualitative synthesis for this systematic review. The study selection process is detailed in Figure 1, which presents the PRISMA flow diagram [20, 23].

Figure 1: PRISMA Diagram

Study Characteristics

The 10 included studies were conducted between 2018 and 2022 and represent diverse geographic regions, including the United States, Japan, the Netherlands, China, and multinational cohorts. Sample sizes across studies ranged from 156 to 951 subjects, with studies enrolling between 300 and 500 subjects. Studies utilized a prospective or cross-sectional design, while others incorporated retrospective analyses.

Populations primarily consisted of older adults with multiple comorbidities, including obesity, hypertension, and atrial fibrillation, consistent with epidemiologic patterns observed in HFpEF [6 , 24]. HFpEF definitions adhered to standard clinical criteria across all studies. Echocardiographic and/or biomarker assessments following established

guidelines were explicitly applied in four studies [6, 21, 25]. Reference standards varied: expert clinical adjudication was used in two studies [26]; invasive hemodynamic testing was employed in two studies [9]; and trial enrollment criteria were used in one study. Table 2 summarizes key characteristics of the included studies, including country, reference standards, and principal findings related to diagnostic discordance. Contextual contributors to discordance, such as comorbidity burden, geographic practice variation, and diagnostic resource availability, are also summarized in Table 2.

Table 2: Study Characteristics, Contextual Factors & Discordance

Study	Country	Sample Size	Reference Standard	Key Findings	Reported Contributors to Discordance
Churchill et al. (2021) [9]	USA	156	Invasive Hemodynamic Testing	~31% discordance; phenotypic variation evident	Intermediate phenotypes; echocardiographic complexity
Selvaraj et al. (2020) [10]	USA	641	ARIC cohort; classification overlap	28% discordance; 4% concordance for high risk	Score weighting; comorbidity profiles
Sanders-van Wijk et al. (2020) [13]	Netherlands	363	Expert Clinical Diagnosis	41% discordance; tool disagreement prominent	Imaging resource variability; threshold definitions

Reddy et al. (2022) [14]	USA / Multinational	736	Multicenter Registry	Higher AUC for H2FPEF; discordance observed	Population heterogeneity; care setting differences
Parcha et al. (2021) [15]	USA / Netherlands	951	TOPCAT / RELAX / ARIC trials	Sensitivity high, specificity differed by score	Trial cohort structure; diagnostic thresholding
Tada et al. (2021) [27]	Japan	234	Clinical diagnosis with echocardiography	Algorithm disagreement noted in diagnostic accuracy	Differences in echo interpretation and thresholds
Sun et al. (2021) [28]	China	358	Clinical algorithm comparison	Score thresholds impacted classification	Tool weighting differences; borderline cases
Egashira et al. (2022) [29]	Japan	502	ESC HFpEF Guidelines	Moderate AUC; scoring cutoff sensitivity	ESC-guided diagnosis; natriuretic peptide thresholds
Amanai et al. (2022) [30]	Japan	187	VO_2 correlation	Scores diverged; VO_2 correlated with score output	Functional vs structural/biomarker inputs

Sueta et al. (2019) [31]	Japan	404	Clinical Evaluation	Context-driven performance variation noted	AF and elderly influence H2FPEF weighting

Note: All studies evaluated H2FPEF and HFA-PEFF algorithms in the same subject populations.

Reporting Study Quality Assessment

Of the 10 included studies, the majority were judged as having low risk of bias across all four QUADAS-2 domains. The QUADAS-2 tool was selected for methodological quality assessment due to its specific design for diagnostic accuracy studies, widespread acceptance in systematic reviews, and ability to evaluate both risk of bias and applicability concerns across key domains. Unlike generic quality assessment tools, QUADAS-2 addresses the unique challenges in diagnostic research, including concerns about index test application, reference standard validity, and patient spectrum.

Three studies were rated as having some concerns in the reference standard domain, primarily due to a lack of clarity regarding blinding or the criteria used to adjudicate HFpEF diagnosis. This highlights the challenge of establishing definitive reference standards in HFpEF diagnosis, where gold standard invasive hemodynamic testing was used in only two studies. Studies with "some concerns" typically relied on clinical diagnosis or guideline-based criteria as reference standards, which introduces potential circular reasoning when evaluating diagnostic algorithms that incorporate similar parameters.

The flow and timing domains were generally well reported, with most studies providing clear sequences of index test application and reference standard determination. In contrast, the index test and patient selection domains showed consistently low risk, with studies clearly describing pre-specified applications of the H2FPEF and HFA-PEFF algorithms. Table 3 presents a detailed summary of the domain-level risk of bias assessment for all included studies.

Table 3: QUADAS-2 Risk of Bias Assessment Summary

Domain	Low Risk	Some Concerns	High Risk
Patient Selection	10/10 (100%)	0/10 (0%)	0/10 (0%)

Index Test	10/10 (100%)	0/10 (0%)	0/10 (0%)
Reference Standard	7/10 (70%)	3/10 (30%)	0/10 (0%)
Flow and Timing	10/10 (100%)	0/10 (0%)	0/10 (0%)

Note: QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies - 2) is a standardized tool for evaluating the methodological quality of diagnostic accuracy studies [22]. The tool assesses the risk of bias across four key domains without generating an overall summary score. The three studies with "some concerns" in the reference standard domain used clinical diagnosis or guideline-based criteria rather than invasive hemodynamic testing.

Discordance Findings Between H2FPEF and HFA-PEFF Algorithms

Discordance between the H2FPEF and HFA-PEFF algorithms was a prominent finding across the included studies. When both algorithms were applied to the same patient cohort, the reported discordance rates ranged from 28% in Selvaraj et al. [10] to 41% in Sanders-van Wijk et al. [13]. Churchill et al. [9] identified substantial variation in patient classification, particularly among those with intermediate pre-test probability, with approximately 31% discordance noted.

Patterns of discordance were associated with specific patient and system-level characteristics. For instance, Churchill et al. [9] noted that subjects with atrial fibrillation or advanced age were more likely to be classified as high-probability HFpEF using the H2FPEF score, which places greater weight on clinical comorbidities. Conversely, subjects lacking definitive imaging or biomarker findings—i.e., intermediate phenotypes—were more variably classified by the HFA-PEFF algorithm. Sanders-van Wijk et al. [13] attributed their 41% discordance rate to imaging availability and adjudication standards ($p = 0.009$). Similarly, Reddy et al. [14] reported that diagnostic agreement varied internationally due to differences in comorbidity burden and resource availability, with statistically significant differences in algorithm performance ($p < 0.001$).

Comparative Diagnostic Performance of H2FPEF and HFA-PEFF Algorithms

Across the included studies, the H2FPEF score generally demonstrated higher sensitivity, while the HFA-PEFF algorithm showed different patterns in specificity. In comparison by Tada et al. [27], H2FPEF demonstrated superior diagnostic accuracy with an AUC of 0.89

compared to HFA-PEFF's AUC of 0.82 (p=0.004). Similarly, Reddy et al. [14] found that H2FPEF had significantly greater area under the curve (0.845, 95% CI, 0.810-0.875) compared with the HFA-PEFF score (0.710, 95% CI, 0.659-0.736), with a statistically significant difference in AUC of 0.134 (p<0.001).

Sanders-van Wijk et al. [13] reported that HFA-PEFF had higher AUC (0.88) compared to H2FPEF (0.77) for their specific cohort (p=0.009), representing an important exception to the overall trend. Sun et al. [28] found that the HFA-PEFF score effectively predicted all-cause mortality (AUC: 0.726, 95% CI: 0.651-0.800, p=0.039), while Egashira et al. [29] reported moderate predictive value for HFA-PEFF (AUC: 0.633, 95% CI: 0.574-0.692, p<0.001) for future HF-related events.

For functional metrics, Amanai et al. [30] found that H2FPEF demonstrated superior feasibility and predictive ability for reduced aerobic capacity compared to HFA-PEFF (AUC: 0.71 vs. 0.61 for predicting peak VO_2, p=0.10). However, this difference did not reach statistical significance. Sueta et al. [31] demonstrated that the H2FPEF score significantly predicted cardiovascular and HF-related events in HFpEF patients (AUC: 0.626–0.680, p<0.001).

These findings confirm that the differential weighting of input variable clinical comorbidities for H2FPEF vs. imaging and biomarkers for HFA-PEFF contributed to the statistically significant discordant classifications observed across multiple studies. Tables 4 and 5 provide a comprehensive comparison of diagnostic and prognostic performance metrics between the H2FPEF and HFA-PEFF algorithms across all included studies.

Table 4: Diagnostic Performance Metrics of H2FPEF and HFA-PEFF Algorithms

Study	H2FPEF Performance	HFA-PEFF Performance	Statistical Comparison
Churchill et al. (2021) [9]	Good overall performance; 28% of low-risk misclassified	Good algorithm performance; 25% of low-risk misclassified	Both had limitations with resting measures; 31% discordance; p-value: NR
Selvaraj et al. (2020) [10]	AUC: NR; Specificity: High for score ≥6; Sensitivity: NR	AUC: NR; Specificity: Similar for score ≥5; Sensitivity: NR	28% discordance in classification; 4% concordance for high risk; p-value: NR

Sanders-van Wijk et al. (2020) [13]	AUC: 0.77; Sensitivity: 70%; Specificity: 82.5%	AUC: 0.88; Sensitivity: 52.7%; Specificity: 90.5%	41% discordance in classification; HFA-PEFF had higher AUC; p=0.009
Reddy et al. (2022) [14]	AUC: 0.845 (95% CI, 0.810-0.875); Higher sensitivity; FNR: 25%	AUC: 0.710 (95% CI, 0.659-0.736); Lower sensitivity; FNR: 55%	H2FPEF superior performance; Difference in AUC: 0.134; p<0.001
Parcha et al. (2021) [15]	Sensitivity: 99.6%; Specificity: 95.6%; PPV: 90.4%; NPV: 99.8%	Sensitivity: 99.5%; Specificity: 82.8%; PPV: 79.9%; NPV: 95.7%	Both highly sensitive; H2FPEF higher specificity & PPV; p-value: NR
Tada et al. (2021) [27]	AUC: 0.89; Specificity: 97% (score 6-9); Sensitivity: 97% (score 0-1)	AUC: 0.82; Specificity: 84% (score 5-6); Sensitivity: 99% (score 0-1)	H2FPEF superior diagnostic accuracy; p=0.004; HFA-PEFF functional sub-score: AUC 0.54
Amanai et al. (2022) [30]	100% calculable; AUC: 0.71 for peak VO_2; Associated with exercise intolerance	88% calculable (missing peptide data); AUC: 0.61 for peak VO_2; Less predictive of exercise capacity	H2FPEF superior feasibility; p=0.10 (non-significant)

Abbreviations: AUC = Area Under the Curve; CI = Confidence Interval; CV = Cardiovascular; FNR = False Negative Rate; HF = Heart Failure; HR = Hazard Ratio; NPV = Negative Predictive Value; NR = Not Reported; PPV = Positive Predictive Value; VO_2 = Oxygen Consumption

Table 5: Prognostic Value of H2FPEF and HFA-PEFF Algorithms

Study	H2FPEF Prognostic Findings	HFA-PEFF Prognostic Findings	Comparative Analysis

Parcha et al. (2021) [15]	No significant prognostic association in TOPCAT	Each 1-point increase: 26% higher hazard for adverse CV events	HFA-PEFF showed stronger prognostic trends; 50% of participants reclassified when applying both scores; p-value: NR
Sun et al. (2021) [28]	Not primary focus of study	AUC: 0.726 (95% CI: 0.651-0.800); Cut-off of 3.5: sensitivity 78.3%, specificity 54.8%; HR: 1.314 for mortality	HFA-PEFF predicted all-cause mortality; Scores 5-6 points: 19.4% increased mortality; p=0.039
Egashira et al. (2022) [29]	Not primary focus of study	AUC: 0.633 (95% CI: 0.574-0.692); Cutoff score of 4.5: sensitivity 57.8%, specificity 67.4%; Independent predictor of HF events	HFA-PEFF predicted future HF events; HR: 1.65; 95% CI: 1.11-2.50; p<0.001
Sueta et al. (2019) [31]	AUC: 0.626 for CV events; AUC: 0.680 for HF events; Cutoff score: 3.5; HR: 1.179-1.288	Not primary focus of study	H2FPEF predicted CV and HF events; Higher scores associated with higher event rates; p<0.001

Abbreviations: AUC = Area Under the Curve; CI = Confidence Interval; CV = Cardiovascular; FNR = False Negative Rate; HF = Heart Failure; HR = Hazard Ratio; NPV = Negative Predictive Value; NR = Not Reported; PPV = Positive Predictive Value; VO

Discussion

This systematic review synthesized evidence regarding discordance between the H2FPEF and HFA-PEFF algorithms in subjects with suspected HFpEF. Across the 10 included

studies, discordance rates ranged from 28% to 41%, highlighting substantial inconsistencies in patient classification between the two clinical algorithms [8, 9, 13].

Although both tools have been integrated into clinical research and referenced in guideline documents, their frequency of use in routine clinical practice remains poorly documented. Available evidence suggests that H2FPEF may be more commonly used in U.S.-based outpatient settings, while HFA-PEFF is more frequently applied in European specialty centers. As noted earlier, real-world adoption of these tools varies across regions and clinical settings, further complicating their comparative evaluation. While most studies had low risk of bias, variability in study design, diagnostic thresholds, and reference standards introduces heterogeneity that must be considered when interpreting these findings.

Recent guidelines and expert consensus emphasize the growing burden of HFpEF and the urgent need for improved diagnostic strategies tailored to diverse patient populations [21]. Furthermore, phenotypic complexity in HFpEF presentation continues to challenge clinicians, necessitating more flexible and comprehensive diagnostic frameworks [27, 29]. Several important patterns emerged from our systematic review. The H2FPEF score generally demonstrated higher sensitivity, often classifying a broader range of subjects as having probable HFpEF. In contrast, the HFA-PEFF algorithm demonstrated higher specificity, frequently requiring more definitive echocardiographic or biomarker evidence for a positive classification. These findings align with prior validation studies suggesting that while the H2FPEF score may better capture cases early in the diagnostic trajectory, the HFA-PEFF algorithm may offer greater accuracy for confirmed diagnoses in more advanced stages [27, 28].

Discordance appeared particularly pronounced among subgroups characterized by obesity, atrial fibrillation, or older age [24, 31]. These subject characteristics are heavily weighted in the H2FPEF score. Still, they may not be as strongly emphasized within the HFA-PEFF framework, which focuses on structural heart disease and diastolic dysfunction parameters [8]. Such differences suggest that scoring criteria performance may be context-dependent, depending on subject demographics and comorbidities.

Social Determinants of Health (SDoH)

Importantly, no included studies systematically incorporated SDoH into their diagnostic frameworks or discordance analyses. This omission represents a critical gap. Extensive literature demonstrates that socioeconomic status, race and ethnicity, access to care, and geographic region materially impact diagnostic access, disease presentation, and

treatment outcomes in heart failure broadly [16 -18]. Given the reliance of both the H2FPEF and HFA-PEFF scores on access to specialized testing (e.g., echocardiography, natriuretic peptide assays), subjects from marginalized populations may be disproportionately affected by underdiagnosis or misclassification when these tools are applied without contextual adjustment [16 - 18].

Emerging consensus highlights the urgent need to integrate SDoH considerations into cardiovascular diagnostics to promote health equity [32]. Advances in explainable artificial intelligence (AI) offer promising pathways to develop future HFpEF diagnostic models that incorporate both clinical and social data [33, 34]. Such models could stratify subjects more equitably, improve diagnostic accuracy, and mitigate underdiagnosis among historically marginalized populations.

Building upon the findings of our prior systematic review evaluating the diagnostic accuracy of the H2FPEF and HFA-PEFF scores, this focused secondary analysis highlights critical limitations in current HFpEF diagnostic tools and underscores the necessity of innovation through social-contextualized and AI-supported models.

Taken together, the findings of this review highlight the limitations of current rule-based diagnostic scoring criteria when applied across heterogeneous populations. While both the H2FPEF and HFA-PEFF algorithms provide valuable structure to the evaluation of suspected HFpEF, neither fully accounts for clinical, demographic, and social variability that influences disease manifestation and diagnostic access.

Implications for Practice and Future Research

The findings of this systematic review have important implications for both clinical practice and the future development of diagnostic tools for HFpEF. The substantial discordance observed between the H2FPEF and HFA-PEFF algorithms underscores the limitations of current rule-based diagnostic scoring criteria when applied across diverse clinical populations [13, 14]. These tools, while valuable, rely heavily on clinical, imaging, and biomarker variables that may not fully capture the complexity and heterogeneity of HFpEF presentations in real-world settings [27].

In clinical practice, the observed discordance suggests that reliance on a single diagnostic scoring criterion may lead to missed or delayed diagnoses, particularly in subjects whose clinical profiles do not neatly align with traditional HFpEF phenotypes. Clinicians should be cautious when interpreting score results in isolation and should consider integrating

multiple sources of information, including subject-specific factors not captured by existing algorithms [21, 32].

Critically, this review highlights a major gap in the current diagnostic landscape: the absence of SDoH integration into HFpEF diagnostic scoring criteria. Variables such as socioeconomic status, race and ethnicity, educational attainment, insurance coverage, and geographic access to care have well-established impacts on diagnostic access, disease progression, and health outcomes in heart failure broadly [16 - 18]. Yet, none of the studies included in this review incorporated SDoH into their diagnostic frameworks or considered social context as a factor influencing algorithm performance.

The integration of SDoH into future HFpEF diagnostic models offers a promising avenue to enhance diagnostic equity. Incorporating social risk factors could help adjust predictive probabilities in ways that account for disparities in access to echocardiography, biomarker testing, and specialist evaluation. Moreover, integrating SDoH could reduce the risk of underdiagnosis in traditionally marginalized populations who may present with atypical or less well-characterized HFpEF phenotypes.

Emerging technologies, particularly explainable artificial intelligence (AI), offer a promising pathway to address diagnostic gaps and enhance equity in heart failure care [33, 34]. Future research should prioritize the development and validation of such integrated models, ensuring that they are trained on diverse, representative populations and that their predictions are interpretable to clinicians. By moving beyond purely biologic frameworks and incorporating social context, next-generation HFpEF diagnostic tools have the potential to improve diagnostic precision, reduce disparities, and promote more equitable cardiovascular care.

Limitations

This systematic review has several important limitations that should be acknowledged. First, although a comprehensive search strategy was employed across multiple databases, there remains the possibility of publication bias. Studies reporting significant discordance between diagnostic algorithms may have been more likely to be published, potentially overestimating the true rate of discordance.

Second, there was considerable heterogeneity across the included studies in terms of study design, subject populations, diagnostic thresholds, and reference standards. Some studies used invasive hemodynamic testing as the diagnostic gold standard, while others relied on expert clinical adjudication or guideline-based classifications. This variability may

have influenced the application of scoring criteria, as well as reported outcomes, diagnostic accuracy, and discordance rates—ultimately limiting the direct comparability of findings across studies.

Third, many of the included studies were conducted in single-center or regional settings, often within specialized heart failure clinics. As a result, findings may not be generalizable to broader, community-based populations, where access to specialized diagnostics may differ substantially.

Fourth, not all studies explicitly quantified discordance rates between the H2FPEF and HFA-PEFF scores. In several cases, discordance was inferred based on differences in reported sensitivity, specificity, or classification thresholds, which introduces some degree of subjectivity in the synthesis of discordance findings.

Fifth, and importantly, this review was limited by the absence of social determinants of health (SDoH) data within the included studies. As no studies systematically incorporated SDoH variables, it was not possible to assess how social context may have influenced discordance rates or diagnostic access. This gap reinforces the need for future research that integrates clinical, imaging, and social variables into HFpEF diagnostic models.

Finally, while a narrative synthesis approach was appropriate given the variability across studies, a formal meta-analysis of discordance rates was not feasible. Quantitative pooling was limited by inconsistent reporting formats, heterogeneity in definitions of positive test classifications, and a relatively small number of directly comparative studies.

Despite these limitations, this systematic review provides important insights into the limitations of current HFpEF diagnostic algorithms and underscores the need for more comprehensive, equitable approaches to diagnosis.

Conclusion

This systematic review demonstrates that significant discordance exists between the H2FPEF and HFA-PEFF diagnostic algorithms when applied to subjects with suspected HFpEF. Discordance rates ranging from 28% to 41% highlight clinical uncertainty that can arise when relying solely on current rule-based diagnostic tools. Differences in sensitivity, specificity, and weighting of clinical versus imaging parameters contribute to inconsistent subjects' classification across diverse populations.

Future diagnostic models for HFpEF should seek to integrate clinical, imaging, and social variables to improve diagnostic precision and promote health equity. Advances in

explainable artificial intelligence (AI) present an opportunity to develop next-generation tools that are both accurate and contextually aware, addressing the limitations of current diagnostic paradigms to ensure timely, equitable diagnosis and management of HFpEF in diverse real-world settings.

References

1. Logeart D. Heart failure with preserved ejection fraction: new challenges and new hopes. *Presse Med*. 2024;53:104185. doi:10.1016/j.lpm.2023.104185
2. Bozkurt B, Ahmad T, Alexander KM, et al. HF STATS 2024: heart failure epidemiology and outcomes statistics—an updated 2024 report from the Heart Failure Society of America. *J Card Fail*. 2025;31:66-1164. doi:10.1016/j.cardfail.2024.10.428
3. Borlaug BA. The pathophysiology of heart failure with preserved ejection fraction. *Nat Rev Cardiol*. 2014;11:507-515. doi:10.1038/nrcardio.2014.83
4. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: executive summary. *J Am Coll Cardiol*. 2022;79:1757-1780. doi:10.1016/j.jacc.2021.12.011
5. Reddy YNV, Carter RE, Obokata M, Redfield MM, Borlaug BA. A simple, evidence-based approach to help guide diagnosis of HFpEF. *Circulation*. 2018;138:861-870. doi:10.1161/CIRCULATIONAHA.118.034646
6. Savarese G, Lund LH. Global public health burden of heart failure. *Card Fail Rev*. 2017;3:7-11. doi:10.15420/cfr.2016:25:2
7. Shah RU, Tsai V, Klein L, Heidenreich PA. Characteristics and outcomes of elderly patients after first HF hospitalization. *Circ Heart Fail*. 2011;4:301-307. doi:10.1161/CIRCHEARTFAILURE.110.959833
8. Pieske B, Tschöpe C, de Boer RA, et al. How to diagnose heart failure with preserved ejection fraction: the HFA-PEFF diagnostic algorithm. *Eur Heart J*. 2019;40:3297-3317. doi:10.1093/eurheartj/ehz641
9. Churchill TW, Li SX, Curreri L, et al. Evaluation of 2 existing diagnostic scores for heart failure with preserved ejection fraction against a comprehensively phenotyped cohort. *Circulation*. 2021;143:289-291. doi:10.1161/CIRCULATIONAHA.120.051299
10. Selvaraj S, Myhre PL, Vaduganathan M, et al. Application of diagnostic algorithms for heart failure with preserved ejection fraction to the community. *JACC Heart Fail*. 2020;8:640-653. doi:10.1016/j.jchf.2020.03.013

11. Barandiarán Aizpurua A, Sanders-van Wijk S, Brunner-La Rocca HP, et al. Validation of the HFA-PEFF score. *Eur J Heart Fail*. 2020;22:413-421. doi:10.1002/ejhf.1614
12. Faxen UL, Venkateshvaran A, Shah SJ, et al. Generalizability of HFA-PEFF and H2FPEF diagnostic algorithms and associations with heart failure indices and proteomic biomarkers: insights from PROMIS-HFpEF. *J Card Fail*. 2021;27:756-765. doi:10.1016/j.cardfail.2021.02.003
13. Sanders-van Wijk S, Barandiarán Aizpurua A, Brunner-La Rocca HP, et al. Discrepancies between H2FPEF and HFA-PEFF scores in diagnosing HFpEF. *Eur J Heart Fail*. 2020;22:838-840. doi:10.1002/ejhf.1700
14. Reddy YNV, Rikhi A, Obokata M, et al. Application of diagnostic algorithms for HFpEF: a comparative analysis. *JAMA Cardiol*. 2022;7:891-899. doi:10.1001/jamacardio.2022.2360
15. Parcha V, Malla G, Kalra R, et al. Diagnostic and prognostic implications of HFpEF scoring systems. *ESC Heart Fail*. 2021;8:2089-2102. doi:10.1002/ehf2.13319
16. Havranek EP, Mujahid MS, Barr DA, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation*. 2015;132:873-898. doi:10.1161/CIR.0000000000000228
17. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. 2017;389:1453-1463. doi:10.1016/S0140-6736(17)30569-X
18. Williams DR, Mohammed SA. Racism and health I: pathways and scientific evidence. *Am Behav Sci*. 2013;57:1152-1173. doi:10.1177/0002764213487340
19. International Prospective Register of Systematic Reviews (PROSPERO). CRD42017075680. *Title of your review registration*. PROSPERO. Published Month Day, Year. Accessed June 26, 2025. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42017075680
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
21. McDonagh TA, Metra M, Adamo M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2021;42:3599-3726. doi:10.1093/eurheartj/ehab368
22. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529-536. doi:10.7326/0003-4819-155-8-201110180-00009
23. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535. doi:10.1136/bmj.b2535

24. Kitzman DW, Shah SJ. The HFpEF obesity phenotype: the elephant in the room. *J Am Coll Cardiol*. 2016;68:200-203. doi:10.1016/j.jacc.2016.05.011
25. Rywik TM, Doryńska A, Wiśniewska A, et al. Epidemiology and clinical characteristics of hospitalized patients with heart failure with reduced, mildly reduced, and preserved ejection fraction. *Pol Arch Intern Med*. 2022;132:16227. doi:10.20452/pamw.16227
26. Owan TE, Hodge DO, Herges RM, et al. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *N Engl J Med*. 2006;355:251-259. doi:10.1056/NEJMoa052256
27. Tada T, Hasegawa K, Nakata M, et al. Prognostic implications of the H2FPEF score in heart failure with preserved ejection fraction. *Circ J*. 2021;85:306-313. doi:10.1253/circj.CJ-20-0846
28. Sun Y, Si J, Li J, et al. Predictive value of HFA-PEFF score in patients with heart failure with preserved ejection fraction. *Front Cardiovasc Med*. 2021;8:656536. doi:10.3389/fcvm.2021.656536
29. Egashira S, Imamura T, Nakano H, et al. The prognostic capability of the HFA-PEFF diagnostic algorithm for heart failure with preserved ejection fraction. *Int Heart J*. 2022;63:487-494. doi:10.2169/ihj.21-779
30. Amanai S, Harada T, Kagami K, et al. The H2FPEF and HFA-PEFF algorithms for predicting exercise intolerance and abnormal hemodynamics in heart failure with preserved ejection fraction. *Sci Rep*. 2022;12:13. doi:10.1038/s41598-021-03974-6
31. Sueta D, Yamamoto E, Nishihara T, et al. H2FPEF score as a prognostic value in HFpEF patients. *Am J Hypertens*. 2019;32:1082-1090. doi:10.1093/ajh/hpz138
32. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*. 2015;131:269-279. doi:10.1161/CIRCULATIONAHA.114.010637
33. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56. doi:10.1038/s41591-018-0300-7
34. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347-1359. doi:10.1056/NEJMra1814259

Appendix I – Full Search Strategy

https://www.crd.york.ac.uk/PROSPEROFILES/2827b4dcd0d3205a624970b3bd0773ec.pdf

$_2$ = Oxygen Consumption