Reproducibility of Echocardiographic Measurements of Left Ventricular Systolic Function: A Systematic Review and Meta-Analysis Comparing Artificial Intelligence and Clinician Estimates.

Rebecca Roberts[1], Leigh Sanyaolu[1], Christina Sam[1], Daniel Farewell[1], Adrian Edwards[1] (edwardsag@cardiff.ac.uk), Rhodri H Davies[2,3]

[1] Division of Population Medicine, Cardiff University, 3rd floor, Neuadd Meirionnydd, Heath Park, Cardiff, CF14 4XN

[2] Institute of Cardiovascular Medicine, University College London

[3] University Hospital of Wales, Cardiff, UK

Abstract

Background

Echocardiography underpins diagnosis and management of cardiovascular disease, yet measurement variability can influence treatment decisions. Artificial Intelligence (AI) may standardise interpretation, but its reproducibility and clinical impact require systematic evaluation.

Objective

To compare the reproducibility of AI-derived and clinician-derived measurements of left ventricular (LV) systolic function, specifically Global Longitudinal Strain (GLS) and Ejection Fraction (EF), in adults.

Methods

We searched Medline, Embase, Web of Science and CENTRAL from inception to May 2025 for peer-reviewed studies assessing reproducibility of AI-derived EF and/or GLS from two-dimensional (2D) or three-dimensional (3D) transthoracic echocardiography. Reporting quality was assessed with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM). Random-effects meta-analyses of Intraclass Correlation Coefficients (ICCs) and Bland-Altman plots compared reproducibility of AI- and clinician-derived measures.

Results

Nineteen studies (17,984 participants; mean age 59 ± 8 years, 52.8% male) were included. Mean CLAIM adherence was 72.9%. Pooled ICCs demonstrated high reproducibility for both AI- and clinician-derived EF and GLS. Bland-Altman analyses showed limits of agreement of –13.4% to +12.7% for 2D EF and –4.3% to +2.3% for 2D GLS. 3D EF was slightly better, showing pooled limits of agreement of 11.26 to 12.61%. The pooled mean absolute differences (MAD) were 5.17% for 2D EF, 5.27% for 3D EF and 1.32% for 2D GLS.

Conclusion

AI-derived GLS and 3D EF achieve reproducibility comparable to, or exceeding, clinicians' estimates. However, limits of agreement between clinician and AI estimates are sufficiently wide that reclassification is possible around key thresholds which could affect patient management decisions. Large-scale, real-world validation remains essential to confirm generalisability.

Keywords

Artificial Intelligence, Echocardiography, Reproducibility, Ejection Fraction, Global Longitudinal Strain, meta-analysis

1. Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, responsible for 18 million deaths per year (1). By 2050, the global burden of CVD is expected to increase substantially. It is estimated that the prevalence of CVD will increase by 90% and result in a 55% increase in Disability-Adjusted Life Years and 73% increase in mortality (2). The projections place increasing demand on diagnostic imaging services for CVD, including echocardiography.

Whilst it is a cornerstone of clinical cardiology and integral to decision-making, echocardiography has limitations. Manual or semi-automatic image analysis is prone to intra-observer and inter-observer variability and is time-consuming (3-6). Additionally, it requires a highly skilled workforce. Globally, pressure on echocardiography services is rising, with national-level data from the UK revealing that an estimated 1.7 million echocardiography scans are requested annually, exceeding current service capacity (7). There is, therefore, an increasing interest in the application of Artificial Intelligence (AI) in this field. AI has shown promise in echocardiography including enhanced diagnostic accuracy, operational efficiency, cost-effectiveness, and improved reproducibility (9). By reducing the variability introduced by differences in operator expertise, patient anatomy and equipment, AI may help improve the consistency of echocardiographic measurements (8).

However, whilst AI may offer a more consistent and reproducible alternative to manual measurements, its integration into clinical practice remains challenging. Barriers include clinician resistance, ethical concerns and its compatibility with existing clinical workflow (8, 9). Thus, the aim of this systematic review is to determine how reproducible echocardiographic measurements of left ventricular function obtained using AI are compared with those calculated by clinicians.

2. Methods

2.1 Study design

This was a systematic review and meta-analysis assessing the reproducibility of echocardiographic measurements of left ventricular function obtained using AI compared to those made by clinicians. All aspects of the review were conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 reporting guideline (10). The protocol for this systematic review is published on PROSPERO (ID: CRD42023477388) (11). As this study involves secondary analysis of published data, ethical approval was not required. All data used were de-identified and included studies with ethical approval from their respective institutions.

2.2 Study Selection

We searched Medline via Ovid using Medical Subject Headings (MeSH terms) combined with Boolean operators, “AND”, “OR”. To reflect the full evolution of AI in echocardiography, we did not apply publication date restrictions and searched databases from inception.

We co-developed our search with an information scientist using search terms related to Artificial Intelligence, echocardiography, and reproducibility, also utilising the search strategies of existing Cochrane reviews and an AI search filter for Medline (12) (13).

Our search (Appendix A) was developed in Medline and retrieved records from inception to the 14th of May 2025. This search was adapted to three additional electronic databases, Embase via Ovid, Web of Science and CENTRAL to capture relevant literature.

Our pilot search identified a substantial number of relevant studies. Only peer-reviewed publications were included. Preprints were excluded to ensure that all included studies had undergone peer review, maintaining methodological and reporting quality, and minimising the risk of including unvalidated findings that could undermine the reliability of our conclusions.

After removing duplicate records using EndNote (Clarivate) and Rayyan (Qatar Computing Research Institute, QCRI), all remaining records underwent title and abstract screening. RR conducted the initial screening based on the predefined eligibility criteria (Table 1). A second reviewer (CS) independently screened a random 10% sample to ensure fairness in inclusion.

The inter-rater reliability, measured using Cohen's kappa (κ = 0.83), indicated 'almost perfect' agreement, so further dual screening was not performed (14).

Disparities between reviewers (CS and RR) were discussed and a consensus was reached for all conflicts and third-party arbitration was not required. At both stages, all excluded studies and reasons for exclusion were documented on the PRISMA 2020 flow diagram (Figure 1). Full-text screening was performed by RR for potentially eligible studies, in consultation with RD, an imaging cardiologist with expertise in AI, to ensure clinical eligibility criteria were met.

Table 1: Eligibility criteria.

Inclusion criteria	**Exclusion criteria**
Papers available in English.	Grey literature.
Peer-reviewed studies. Primary research.	Reviews, systematic reviews, meta-analyses, editorials, letters, conference posters, case reports, opinions, book chapters, letters, case series, commentaries, conference papers or

	posters, proceedings, dissertations, and thesis submissions.
Reporting the reproducibility of echocardiographic measurements of left ventricular function using AI.	Not reporting the reproducibility of echocardiographic measurements of left ventricular function using AI.
2D- or 3D-transthoracic echocardiography.	Studies using other types of imaging techniques.
Comparator group is clinicians or specialist physiologists.	No comparator group, or comparator group not involving clinicians or specialist physiologists.
Adult human participants (18 and over) without congenital heart disease.	Participants under 18 or those with congenital heart disease.
Reporting of at least one reproducibility metric, including the Coefficient of Variation (CoV), Cohen's κ, Bland-Altman limits of agreement (LoA) and Intraclass Correlation Coefficient (ICC).	No reporting of reproducibility statistics.

2.4 Data extraction

A summary of the extracted information is presented in Table 2. Subsequently, data were assessed for suitability for meta-analysis, depending on the level of heterogeneity across studies, principally the type of reproducibility metrics presented.

Table 2: Data extraction.

Study characteristics	Title, author, publication year, journal, nationality.
Methodology	Study design, study population echocardiographic modality, comparator methods.
AI usage	Type of AI model (DL, ML), validation.
Reproducibility metrics	ICC, Bland-Altman Limits of Agreement, CoV, Cohen's κ.
Strengths and limitations	As acknowledged by the study.

Conclusion	Overall conclusion, study funding, other strengths and limitations, conflict of interest.

Abbreviations: DL = Deep Learning; ML = Machine Learning; CoV = Coefficient of Variation; ICC = Intraclass Correlation Coefficient.

2.5 Quality assessment

We used the 42-item Checklist for Artificial Intelligence in Medical Imaging (CLAIM) to assess the quality of evidence (15). Whilst CLAIM is primarily a reporting tool rather than an appraisal tool, it offers a structured framework for evaluating the transparency and methodological rigor of AI-based medical imaging studies in the absence of a widely accepted AI risk of bias tool specific to reproducibility studies.

Quality assessment evaluated each study against the key CLAIM domains:

- Study design and patient selection.
- AI model development and validation.
- Comparator and ground truth.
- Performance metrics and reproducibility.
- Clinical relevance and generalisability.

The mean and standard deviation (SD) of CLAIM scores was also calculated to summarise the overall reporting quality.

2.6 Data synthesis

Reproducibility refers to the consistency of repeated measurements on the same subject under varying conditions and is essential for clinical reliability (8). It can be assessed using metrics such as the Coefficient of Variance (CoV), Cohen's κ, Bland-Altman limits of agreement (LoA), and the Intraclass Correlation Coefficient (ICC) (8).

The ICCs and Bland-Altman analyses were consistently reported across studies enabling meta-analysis. Due to methodological and clinical differences across studies, including variations in AI models, study populations, and comparator groups, a random-effects model was used to allow for variation in the underlying effect sizes (16).

To conduct the ICC meta-analysis, we first estimated the variance using the formula by Bhat and Beretvas (17):

$$\text{ICC Variance} = (2 \times (1 - \text{ICC})^2/(\text{n} - 1)) \qquad \textit{where } n \textit{ = sample size}$$

Anticipating significant heterogeneity between studies, we conducted separate random-effects meta-analyses using the standard error to obtain pooled inter-observer, intra-

observer, inter-technique and AI-derived ICCs (18). Inter-observer ICCs reflect agreement between different human observers; intra-observer ICCs measure the consistency of the same observer over time; inter-technique ICCs assess reproducibility between clinician and automated measurements, and AI-derived ICCs reflect the consistency of repeated measurements made by the AI model. The results are summarised in tables within the results section.

Despite the popularity of Bland-Altman plots, meta-analytical methods for synthesising their results are underdeveloped (19). Hence, we followed the framework proposed by Tipton and Shuster to conduct an inverse-variance weighted meta-analysis of the bias (18, 19). This process involved calculating the standard deviation of differences, assessing normality, determining inverse-variance weights, and computing both the weighted bias and the mean weighted bias, along with their standard error and 95% confidence intervals (CI) (19). The full details and equations used are available in Appendix B.

The standard error for the LoA was estimated as (20):

$SE = Sd\ x \sqrt{\frac{1}{n} + \frac{(1.96)^2}{2(n-1)}}$ *where Sd = standard deviation of differences; n = sample size*

Two separate random-effect meta-analyses for the upper and lower LoA were performed to obtain a pooled estimate. The results are displayed graphically, illustrating the individual

study estimates alongside the meta-analysed result, with 95% CI and standard errors for each data point.

All meta-analyses were conducted using the statistical package SPSS (version 29.0.2.0, IBM). Bland-Altman analyses are presented using Microsoft Excel (version 16.87, Microsoft Corp).

In addition to the meta-analyses, we generated graphs to contextualise measurement variability in relation to guideline-defined clinical thresholds for EF and GLS. For each study, Bland-Altman limits of agreement were extracted and entered into Excel to illustrate agreement between clinician and AI estimates at decision-relevant cut-offs. The mean absolute difference (MAD) was calculated from the standard deviation of differences, approximating the average magnitude of error under the assumption of normally distributed differences, using the formula (21):

$$MAD = (2\ x\ Sd) \div \pi$$

We further calculated mean absolute percentage difference (MAPD) by dividing the MAD by the relevant threshold and multiplying by one hundred, thereby expressing error size relative to clinical thresholds.

3. Results

3.1 Study selection

The search identified 5,040 records (Medline via Ovid: 1,044, Embase via Ovid: 2,279, Web of Science: 1,347, CENTRAL: 370). After removal of duplicates (n = 1,918), 3,122 records remained for initial title and abstract screening.

2,952 records did not meet the eligibility criteria, leaving 170 studies for full-text screening. Of these, a further 151 were excluded and nineteen studies were included in our systematic review. A PRISMA 2020 flow diagram illustrating the study selection process is shown in Figure 1.

***Figure 1:** PRISMA flow diagram depicting study selection process, including records from searches in December 2024 and May 2025 (10).

**See separate image file.*

3.2 Study characteristics

Of the nineteen studies, ten assessed the reproducibility of AI-derived EF, four assessed AI-derived GLS, and five evaluated both. Participant demographics were available for eighteen of these studies and enrolled 17, 984 participants (22-37). Of these, 9,487 (53%) were male.

The mean age was 59 ± 8 years. Table 3 provides a summary of study characteristics, including sample size, AI model type, the clinical comparator (e.g., expert clinician, sonographer) and CLAIM score.

Heterogeneity of findings within each random effects model ranged from 45-100% and thus was generally substantial (Appendix C).

3.3 Study Quality

The mean CLAIM score across all studies was 30 ± 4 out of a maximum score of 42 (72.9%), indicating moderate to high adherence with reporting standards. The highest mean percentage completion is seen in title/abstract (100%), introduction (100%), study design (100%) and discussion (100%) domains. The poorest are seen in other information (35.1%) and training (45.6%) domains. A summary of CLAIM domain scores across all included studies can be found in Appendix D.

Table 3: Summary of study participant demographics and CLAIM scores across all included studies.

Measure of LV function	2D/3D TTE	Study	Sample	Study characteristics					AI model	Comparator	CLAIM score
				Male	Female	Age	Arrhythmia	VHD			
LVEF	2D	Asch et al. (2019) (37)	99	62	37	66.0	n.d	n.d	ML algorithm (AutoEF, Bay Labs Inc.)	Three echocardiographers	28
	2D	Kim et al. (2022) (34)	500	251	249	36.2	n.d	0	Three DL algorithms (U-net, Res-U-net, Dense-U-net)	Two sonographers	26
	3D	Medvedofsky et al. (2018), (31)	180	119	61	57.0	29	20	ML, HeartModel (Philips Healthcare)	*"Trained personnel at a highly experienced Core Laboratory (CL)"*	35
	2D	Mor-Avi et al. (2023), (30)	12	n.d	n.d	n.d	n.d	n.d	DL algorithm	Ten echocardiographers	30
	2D	Morbach et al. (2024), (29)	4965	2404	2,561	54.9	n.d	n.d	ML within a federated learning framework	*"Trained and internally certified personnel performed measurements"*	37
	2D/3D	Myhr et al. (2018), (28)	100	38	62	67.0	10	37	AutoEF (2D) 4D Auto LVQ (3D)	One sonographer	31
	2D	Olaisen et al. (2024), (25)	3,282	1865	1417	59.5	322	60	DL based on U-net architecture for segmenting LV endocardium, myocardium & atrium	Dataset 1: three cardiologists Dataset 2: four 'experienced operators' Dataset 3: two cardiologists Dataset 4: sonographer, cardiologist	29
	2D	Sveric et al. (2023), (38)	889	542	347	71.0	223	181	DL algorithm, LVivo Seamless™	Cardiologist	30
	2D	Li et al. (2025), (39)	2461	1329	1132	52.4	n.d	n.d	DL custom modular model based on YOLOX	*"two highly experienced doctors, each*	33

										possessing over 10 years of clinical expertise in cardiac ultrasound…"	
	2D	Lin et al. (2024), (40)	2613	1163	1450	56	n.d	n.d	QHAutoEF, integrating DL and transformers	Multiple senior echocardiographers.	40
GLS	2D	Nyberg et al. (2024), (26)	80	48	32	61.0	n.d	n.d	DL-method based on point tracking.	*"Three experienced observers"*	26
	2D	Rogstadkjernet et al. (2024), (23)	605	372	233	63.4	n.d	143	DL segmentation model, EfficientNetB1	Echocardiographers	34
	2D	Salte et al. (2023), (22)	72	42	30	63.5	n.d	n.d	DL-method	Four echocardiographers	28
	2D	Kuwahara et al. (2024), (32)	94	25	69	69.0	n.d	n.d	U-Net DL for endocardial segmentation, *Caas Qardia 1.1 software*	*"Experienced examiners"*	34
LVEF, GLS	2D	Jang et al. (2024), (36)	632	531	101	59.3	n.d	n.d	3D DL segmentation & motion estimation	Sonographers, echocardiographers	26
	2D	Knackstedt et al. (2015), (33)	255	153	102	50.3	Excluded.	4	ML-algorithm, AutoLV	*"Expert investigator (level 3 training in echocardiography; C.K., A.F., L.B., P.S.)"*	33
	2D	Lafitte et al. (2025), (41)	894	510	384	64.8	n.d	n.d	Deep learning, Us2.ai	*"Operators with three experience levels (nurses, residents and experts)"*	27
	2D	Jiang et al. (2023), (35)	142	2	140	59.0	Excluded.	n.d	DL, Ligence Heart (version 2)	*"BSE-accredited or similarly experienced*	29

										operator manually adjusted the ROI"	
	2D/3D	Myhre et al. (2024), (27)	109	32	77	56.0	Excluded.	n.d	2D: Us2.ai 3D: Heartmodel3D, echo analysis	*"Experienced operator had EACVI transthoracic echocardiography certification or an echocardiography experience of more than 10 years"*	26

Abbreviations; *AI = Artificial Intelligence; EF = Ejection Fraction, GLS = Global Longitudinal Strain; LV = Left ventricle; TTE = Transthoracic Echocardiography; VHD = Valvular Heart Disease; DL = Deep Learning; Machine Learning = Machine Learning; CLAIM = Checklist for Artificial Intelligence in Medical Imaging; n.d = not disclosed.*

3.4 Ejection Fraction (EF)

To contextualise the pooled reproducibility estimates, Supplementary Table 1 presents the raw mean EF values and their standard deviations as reported by clinicians and AI across the studies included in our review. AI-derived EF values were broadly comparable to those of clinicians, with a slight tendency towards lower estimates with similar variability.

ICCs for ejection fraction (EF) were reported in ten studies using 2D TTE, with no data available for 3D TTE. Only one study reported an AI-derived ICC of 0.92 (95% CI: 0.900, 0.936) precluding meta-analysis (37). We undertook meta-analyses of human inter-observer, human intra-observer and inter-technique ICCs for comparison (Table 4). Although AI-specific data were limited, the reported AI ICC (0.92) exceeded these others, suggesting that AI may yield more reliable EF measurements from 2D TTE images than manual methods performed by either the same or different clinicians.

Table 4: Pooled Intraclass Correlation Coefficients (ICCs) with 95% CI for manually derived EF and inter-technique agreement from 2D TTE images.

	Pooled ICC (95% CI)
Manually derived EF inter-observer agreement (27, 33, 34)	0.83 (0.76, 0.91)
Manually derived EF intra-observer agreement (27, 28, 33, 34, 37)	0.88 (0.82, 0.95)
Inter-technique agreement (28, 33, 36-41)	0.85 (0.80, 0.88)

Figure 2 presents the Bland-Altman results depicting agreement between AI and clinicians' EF values from 2D TTE images. Across the included studies, bias values ranged from -5.5% to 4.5% (total range 10.0%). The pooled mean bias was -1.4% (95% CI: -1.35, -1.42), consistent with the raw data in Supplementary Table 1 which shows that AI generally produces lower EF estimates than clinicians. The pooled LoA are wide, ranging from -13.44% (95% CI: -16.19, -10.69) to +12.7% (95% CI: 10.59, 14.82) indicating significant variability in AI-derived EF measurements relative to the clinicians' reference point.

Figure 3 displays the Bland-Altman results showing the agreement between AI- and clinician-derived EF from 3D TTE images. Similarly, a pooled bias of -0.58% (95% CI: -0.72, -0.44) indicates that AI values are marginally lower than clinician derived EF. Notably, the magnitude of the bias in 3D TTE is smaller than that observed with 2D TTE. Variability is also slightly reduced as reflected by marginally narrower pooled LoA for 3D EF, ranging from -11.26% (95% CI: -15.58, -6.95) to 12.61% (95% CI: 6.55, 18.67), indicating a slight improvement in consistency in AI-derived EF from 3D TTE compared to 2D TTE images.

To contextualise these findings against clinical thresholds, we evaluated agreement between clinician- and AI-derived EF from both 2D and 3D TTE images at guideline-defined cut-off points (figure 4). Normal left ventricular systolic function is defined as > 52% in men

and >54% in women, while an EF < 35% represents the threshold for considering cardiac resynchronization therapy (CRT) or primary prevention ICD implantation (42, 43).

For 2D TTE, Figure 4a demonstrates that variability between AI- and clinician-derived EF is considerable, with a MAD of 5.17%. This corresponds to MAPD of 14.8% at the 35% treatment threshold and 9.9% at the 53% "normal function" cut-off. Discrepancies of this magnitude have the potential to alter clinical classification at clinical decision-making thresholds.

For 3D TTE, Figure 4b also shows considerable variability, with a MAD of 5.27%, corresponding to a MAPD of 15.1% at the 35% threshold and 9.9% at the 53% threshold. Although the magnitude of error is almost identical to that observed with 2D TTE, the persistence of variability across both modalities suggests that disagreement between AI and clinician assessment is unlikely a limitation of one imaging approach but rather reflects a systematic challenge in reproducibility.

***Figure 1:** Bland-Altman plot presenting agreement between AI-derived and Clinicians' EF from 2D TTE images. Individual study results and pooled results from meta-analysis shown.

***Figure 2:** Bland-Altman plot presenting agreement between AI-derived and Clinicians' EF from 3D TTE images. Individual study results and pooled results from meta-analysis shown.

***Figure 4:** Uncertainty in AI- versus clinician-derived ejection fraction (EF) measurements around guideline-defined clinical thresholds. (A) 2D TTE studies, (B) 3D TTE studies.

**See separate image files.*

3.5 Global Longitudinal Strain (GLS)

Supplementary table 2 summarises the raw mean GLS values with standard deviations reported in the included studies from 2D TTE images. While broadly aligned with clinician measurements, AI-derived GLS tended to be marginally lower and demonstrated reduced variability.

Table 5 presents the pooled ICC values with 95% CIs for agreement in GLS measurements between AI and clinicians' values from 2D TTE images. Manual intra-observer had an ICC of 0.85 (95% CI: 0.77, 0.93) compared to ICC of 0.81 for AI (95% CI: 0.75, 0.87, both indicating good reproducibility. The overlapping but narrower confidence intervals suggest that AI's consistency may be comparable to or potentially exceed, manual intra-observer interpretation. Manual inter-observer agreement was lower with only moderate reproducibility (ICC = 0.75; 95% CI: 0.68-0.83) highlighting greatest variability between clinicians' assessments. The inter-technique ICC (ICC = 0.77; 95% CI: 0.65, 0.88) indicates moderate agreement; however, the wide confidence interval suggests uncertainty in this value.

Table 5: Pooled ICC (95% CI) for agreement in GLS measurements from 2D TTE, comparing AI-derived and manual methods.

	Pooled ICC (95% CI)
AI-derived GLS agreement (22, 32)	0.81 (0.75, 0.87)
Manually derived GLS inter-observer agreement (22, 26, 32, 33)	0.75 (0.68, 0.83)
Manually derived GLS intra-observer agreement (22, 26, 33)	0.85 (0.77, 0.93)
Inter-technique agreement (26, 27, 35, 36, 41)	0.77 (0.65, 0.88)

Figure 5 presents the Bland-Altman results depicting agreement between AI and clinician-derived GLS from 2D TTE images. Seven studies reported Bland-Altman results and were included in the meta-analysis. The pooled mean bias was -0.80 (95% CI: -0.89, -0.76). Consistent with trends observed in AI-derived EF, AI produces lower estimates of GLS but nonetheless demonstrates excellent agreement with manual measurements. Furthermore, the pooled LoA are narrow and reflect low variability ranging from -4.30% (95% CI: -6.15, -2.28) to +2.30% (95% CI: 1.72, 2.86).

To contextualise these findings against clinical thresholds, we evaluated agreement between clinician- and AI-derived GLS from both 2D TTE image at the threshold for a normal GLS, 20% (Figure 6). We obtained a MAD of 1.32%, corresponding to a MAPD of 6.61%.

***Figure 5:** Bland-Altman plot presenting agreement between AI-derived and Clinicians' GLS from 2D TTE images. Individual study results and pooled results from meta-analysis shown.

***Figure 6:** Agreement between clinician- and AI-derived Global Longitudinal Strain (GLS) values from 2D transthoracic echocardiography (TTE) at guideline-defined clinical thresholds.

**See separate image file.*

4. Discussion

4.1 Principal findings

This systematic review assessed the reproducibility of echocardiographic measurements obtained using AI, compared to those calculated by clinicians. The findings suggest that AI-derived GLS from 2D transthoracic echocardiogram (TTE) images and EF from 3D TTE images are reproducible. Measurement variability, evidenced by Bland-Altman LoA, are reported for AI-derived EF from 2D TTE, AI-derived EF from 3D TTE, and 2D AI-derived GLS, with each modality showing distinct limits of agreement.

AI-derived intraclass correlation coefficients (ICC) for both EF and GLS were high, and each exceeded human inter-observer ICC. While the manual intra-observer ICC for GLS was marginally higher than the AI-derived ICC, overlapping and narrower confidence intervals suggest that AI has the potential to at least replicate the level of consistency expected from a single trained operator. Since AI holds the potential to standardise measurements across clinicians, settings, and institutions this level of reliability is clinically promising.

4.2 Existing literature

These findings complement a growing body of literature reporting that AI can enhance reproducibility in echocardiography by reducing inter-observer variability inherent in manual measurements (5, 44). Nonetheless, concerns remain that AI reproducibility is highly dependent on image quality, a limitation highlighted by several studies included in this review (31, 34, 35, 37, 45). Image quality can be affected by differences in equipment and patient-related factors including body habitus, arrhythmia, and respiratory patterns, all of which may introduce noise and artefacts that reduce algorithmic accuracy and impact reproducibility (46, 47).

Similarly, the improved reproducibility observed in clinician-derived EF using 3D TTE compared to 2D TTE has been attributed to differences in image acquisition. 3D TTE offers more comprehensive volumetric data, reducing geometric assumptions and thereby improving consistency in image alignment and endocardial border detection (48). These same advantages likely contribute to greater reliability in AI-derived EF from 3D datasets as well.

Improvements in reproducibility with AI integration have also been observed in other cardiac imaging modalities, including cardiac MRI and CT, supporting the broader claim that AI can enhance standardisation and reproducibility in cardiac imaging, provided that image quality is adequate (49, 50).

4.3 Clinical Implications

The considerable variation observed in clinicians' 2D EF measurements has potential for profound clinical implications. EF is the gatekeeper for many drugs and devices, and any error in EF measurement has the potential to affect clinical decision-making (51). Figures 4 and 6 which illustrate agreement between clinician and AI EF estimates from 2D and 3D TTE images around clinical decision-making thresholds highlights the risk that measurement variability may result in patient reclassification. For EF, even modest discrepancies could shift an individual across treatment boundaries, with potential consequences for device eligibility or for distinguishing normal from impaired systolic function (42, 43).

AI-derived GLS demonstrated high precision, with narrow LoA, outperforming clinicians' interobserver variability, which range from 5.4% to 8.6% in the literature (52). This higher degree of reproducibility is particularly encouraging given the growing reliance on GLS as a marker for early detection of subclinical myocardial dysfunction, particularly among patients receiving cardiotoxic chemotherapies such as anthracyclines or trastuzumab (53, 54). Given that small, clinically meaningful changes in GLS can influence management, measurement precision is paramount. Greater variability between human observers may mask these subtle changes, misinforming clinical decision-making. In contrast, AI's ability to deliver more consistent measurements and reduce interobserver variability may mitigate these challenges. Other key potential benefits include reducing reliance on a highly specialised workforce for image interpretation, thereby alleviating pressure on

overstretched echocardiography services (4). Taken together, these findings suggest that AI-derived GLS may represent the most immediate avenue for clinical deployment, offering greater reproducibility than EF and reducing risk of misclassification in scenarios where subtle changes matter.

Despite overall moderate to high adherence with the CLAIM reporting tool, the "training" domain had the lowest average score. The training section covers details of the description of training data, labelling and ground truth, preprocessing, dataset splitting, handling of bias and availability of the dataset (15). To ensure equity and safety, future policy must prioritise transparency in training data which will also allow clinicians in other institutions to replicate and externally validate AI tools to assess reproducibility. An additional important implication is the potential for automation bias. There is evidence to suggest that clinicians often lack the technical expertise to critically assess AI, leading to unintentional misuse and overreliance on model outputs (55, 56).

4.4 Current limitations of AI in echocardiography

Despite AI's promise, several limitations must be acknowledged. It is well established that echocardiographic images differ substantially between ultrasound machines and manufacturers, as well as across image analysis software packages (57, 58). While AI algorithms can standardise interpretation to reduce variability introduced during image

analysis, they cannot yet overcome fundamental differences in raw image acquisition between vendors. With several global vendors routinely used in clinical practice, this variability poses a persistent barrier to universal application, as both AI- and clinician-derived measurements can be influenced by variability in acquisition and analysis hardware and software (59, 60). Standardisation of definitions, image acquisition protocols, analysis techniques, and reporting workflows, for example, in speckle-tracking echocardiography, can help mitigate these discrepancies (54). Recent research has explored the use of adversarial AI models to minimise inter-domain variability, while training on federated datasets has shown promise in improving vendor-independence and cross-domain generalisability (29, 61, 62). The broader issue of generalisability across diverse populations and clinical settings is considered further in our Future Research section.

In addition, LVEF and GLS have no definitive "ground truth," so even highly reproducible AI estimates cannot be assumed to be fully accurate. Differences observed between AI and manual (or semi-automatic) measurements therefore reflect variability in the assessment methods rather than necessarily indicating under- or overestimation.

4.4 Future research

Future studies should focus on the external validation of AI models across diverse populations, institutions and vendor equipment. We included both true test–retest studies, where two separate imaging acquisitions were performed in the same patient, and studies

that compared images acquired within a single session but from different cardiac cycles. While both approaches yield different images, test-retest studies are preferable, as they more closely reflect real-world clinical variability, including differences in probe placement, patient positioning, and operator technique. This is particularly relevant for the management of chronic cardiac disease, where serial echocardiographic assessment of LV function, often performed by the different clinicians, are used to monitor long-term disease progression or treatment response in the same patient (63). In this context, evaluating test-retest reliability is essential to determine whether changes in cardiac function reflect true disease progression or are simply due to measurement variability (63).

Furthermore, the consequence of AI training on predominantly male TTE images cannot be ignored and the impact of this needs to be explored. It is imperative to develop methods to mitigate the potential gender bias and promote equity in algorithmic training.

Yet, publishing the results of AI systems is only the first step towards clinical translation. The next stage requires demonstrating improvements in patient outcomes and establishing cost-effectiveness (64). Regulatory approval, prospective clinical validation, and integration into existing workflows remain substantial challenges hence most published AI algorithms are never implemented in practice (65).

To support clinical integration, future research should focus on large-scale, prospective validation of AI models across diverse populations and settings, transparent reporting of training data, mitigation of bias, and demonstration of improvements in patient outcomes and cost-effectiveness. Collaboration between clinicians, industry, and regulators will be essential to ensure safe and effective implementation into routine echocardiographic practice.

4.5 Strengths and limitations

To our knowledge, this is the first systematic review to evaluate and meta-analyse the reproducibility of AI-derived echocardiographic measurements of left ventricular function. The methodology was rigorous, adhering to PRISMA 2020 guidelines, and our search strategy was co-developed with an information scientist. Included studies had moderate to high quality in reporting, with a mean CLAIM score of 31 ± 4 out of 42 (72.9%) - higher than the median score of 26 reported across 421 AI imaging studies from 1997-2024 (66). A further strength lies in the novel application of meta-analytic methods for Bland-Altman analyses using Tipton and Shuster's framework (19).

Several limitations should also be acknowledged. In some studies, the level of experience of the clinical comparator was unclear, potentially introducing comparator bias as reproducibility is known to be influenced by the operator's expertise (41, 67). Evidence

suggests that less experienced operators experience greater variability in measurement interpretation, consistency and diagnostic accuracy (67).

Furthermore, caution is warranted given the heterogeneity within the models (Appendix C), although, this was to be expected given the relatively small number of included studies. As a result, estimates of heterogeneity are likely to be imprecise (68).

Generalisability is a key limitation. Many AI algorithms were developed using single-centre, single-vendor datasets, which may limit their applicability across different clinical settings (28, 35, 38). Although the overall study population was slightly male-predominant (53%), individual studies varied significantly in sex distribution, with the highest discrepancy reflected by a 5:1 male-to-female ratio (36). This raises concerns regarding the applicability of findings to female patients, particularly in light of known sex-based differences in cardiac structure and function that may impact measurement reproducibility (69).

Additionally, a significant number of studies were conducted retrospectively (25, 27, 32, 34-37). Prospective studies including real-time validation are needed to evaluate how these algorithms perform when integrated into routine care.

We were unable to perform subgroup analyses by clinical condition, as very few conditions were specified in the original studies, and those that were reported were poorly defined. This

is an important limitation given that differences in underlying pathology may substantially influence results and contribute to selection bias. This is particularly relevant for AI which typically performs best when it has been trained on similar examples (70). For instance, if a model has not encountered an unusual morphology during training, such as a large apical aneurysm with associated thrombus, it is unlikely to recognise or correctly analyse it. In contrast, a clinician drawing on prior knowledge and experience, would still be able to interpret such findings.

Where possible, EF should be measured quantitatively using Simpson's biplane method, however some studies, relied on visual estimations (51). In studies that presented Bland-Altman plots comparing visual and biplane methods, the biplane measurements demonstrated tighter clustering patterns. Although similar clustering in other studies imply use of the same method, this cannot be conclusively determined.

5. Conclusion

AI-derived GLS, and to a lesser extent 3D EF, demonstrate reproducibility equal to or exceeding clinicians, positioning GLS as the most reliable entry point for clinical AI adoption. Cautious interpretation of 2D EF is warranted due to its greater variability, which may affect reproducibility and lead to inconsistencies in clinical decision-making.

Conflict of Interest

Dr Rhodri H Davies declares share ownership in Mycardium AI.

Acknowledgements

We thank Elizabeth Gillen of Cardiff University for her advice whilst developing the literature search.

References

1. World Health Organization. Cardiovascular diseases Geneva: World Health Organization; [Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.
2. Chong B, Jayabaskaran J, Jauhari SM, Chan SP, Goh R, Kueh MTW, et al. Global burden of cardiovascular diseases: projections from 2025 to 2050. Eur J Prev Cardiol. 2024.
3. Barris B, Karp A, Jacobs M, Frishman WH. Harnessing the Power of AI: A Comprehensive Review of Left Ventricular Ejection Fraction Assessment With Echocardiography. Cardiology in Review. 2024;23:23.
4. Vidal-Perez R, Grapsa J, Bouzas-Mosquera A, Fontes-Carvalho R, Vazquez-Rodriguez JM. Current role and future perspectives of artificial intelligence in echocardiography. World Journal of Cardiology.15(6):284-92.
5. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature.580(7802):252-6.
6. Akerman AP, Porumb M, Scott CG, Beqiri A, Chartsias A, Ryu AJ, et al. Automated Echocardiographic Detection of Heart Failure With Preserved Ejection Fraction Using Artificial Intelligence. JACC Advances.2(6):100452.
7. Colebourn C. State of the art: A roadmap for the national echocardiography team 2023. Future Healthcare Journal.11(1):100129.
8. Bunting KV, Steeds RP, Slater K, Rogers JK, Gkoutos GV, Kotecha D. A Practical Guide to Assess the Reproducibility of Echocardiographic Measurements. Journal of the American Society of Echocardiography.32(12):1505-15.

9. Hua D, Petrina N, Young N, Cho JG, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: A scoping review. Artificial Intelligence in Medicine.147:102698.
10. 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.
11. National Institute for Health Research. PROSPERO Registration: CRD42023477388: University of York; 2024 [Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023477388.
12. Kang C, Lo JE, Zhang H, Ng SM, Lin JC, Scott IU, et al. Artificial intelligence for diagnosing exudative age-related macular degeneration. Cochrane Database of Systematic Reviews. 2024(10).
13. Campbell SMaK, J. Filter to Retrieve Studies Related to Artificial Intelligence from the OVID MEDLINE Database2024 [cited 2025 March 31]. Available from: https://docs.google.com/document/d/1eWyO0jv9_6FYsxyC5LUYwFe9eH_3h83-tPNZ6wmos18/edit#heading=h.qi55eeyvgzy9.
14. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159-74.
15. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artificial Intelligence. 2020;2(2):e200029.
16. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1(2):97-111.
17. Bhat BH, Beretvas SN. Meta-Analytic Pooling of Intraclass Correlation Coefficient Estimates. Multivariate Behav Res. 2022;57(1):156.
18. IBM Corp. IBM SPSS Statistics for Mac, Version 29.0.2.0. Armonk, NY: IBM Corp; 2023.
19. Tipton E, Shuster J. A framework for the meta-analysis of Bland-Altman studies based on a limits of agreement approach. Stat Med. 2017;36(23):3621-35.
20. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135-60.
21. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-51.
22. Salte IM, Østvik A, Olaisen SH, Karlsen S, Dahlslett T, Smistad E, et al. Deep Learning for Improved Precision and Reproducibility of Left Ventricular Strain in Echocardiography: A Test-Retest Study. J Am Soc Echocardiogr. 2023;36(7):788-99.
23. Rogstadkjernet M, Zha SZ, Klæboe LG, Larsen CK, Aalen JM, Scheirlynck E, et al. A deep learning based method for left ventricular strain measurements: repeatability and accuracy compared to experienced echocardiographers. BMC Medical Imaging. 2024;24(1):305.
24. Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, et al. AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE Journal of Biomedical & Health Informatics.24(7):1837-57.

25. Olaisen S, Smistad E, Espeland T, Hu J, Pasdeloup D, Østvik A, et al. Automatic measurements of left ventricular volumes and ejection fraction by artificial intelligence: clinical validation in real time and large databases. European Heart Journal - Cardiovascular Imaging. 2023;25(3):383-95.
26. Nyberg J, Østvik A, Salte IM, Olaisen S, Karlsen S, Dahlslett T, et al. Deep learning improves test–retest reproducibility of regional strain in echocardiography. European Heart Journal - Imaging Methods and Practice. 2024;2(4).
27. Myhre PL, Gaibazzi N, Tuttolomondo D, Sartorio D, Ugolotti PT, Covani M, et al. Concordance of left ventricular volumes and function measurements between two human readers, a fully automated AI algorithm, and the 3D heart model. Front Cardiovasc Med. 2024;11:1400333.
28. Myhr KA, Pedersen FHG, Kristensen CB, Visby L, Hassager C, Mogelvang R. Semi-automated estimation of left ventricular ejection fraction by two-dimensional and three-dimensional echocardiography is feasible, time-efficient, and reproducible. Echocardiography. 2018;35(11):1795-805.
29. Morbach C, Gelbrich G, Schreckenberg M, Hedemann M, Pelin D, Scholz N, et al. Population data–based federated machine learning improves automated echocardiographic quantification of cardiac structure and function: the Automatisierte Vermessung der Echokardiographie project. European Heart Journal - Digital Health. 2023;5(1):77-88.
30. Mor-Avi V, Blitz A, Schreckenberg M, Addetia K, Kebed K, Scalia G, et al. Deep learning assisted measurement of echocardiographic left heart parameters: improvement in interobserver variability and workflow efficiency. The International Journal of Cardiovascular Imaging. 2023;39(12):2507-16.
31. Medvedofsky D, Mor-Avi V, Amzulescu M, Fernández-Golfín C, Hinojar R, Monaghan MJ, et al. Three-dimensional echocardiographic quantification of the left-heart chambers using an automated adaptive analytics algorithm: multicentre validation study. Eur Heart J Cardiovasc Imaging. 2018;19(1):47-58.
32. Kuwahara A, Iwasaki Y, Kobayashi M, Takagi R, Yamada S, Kubo T, et al. Artificial intelligence-derived left ventricular strain in echocardiography in patients treated with chemotherapy. Int J Cardiovasc Imaging. 2024;40(9):1903-10.
33. Knackstedt C, Bekkers SC, Schummers G, Schreckenberg M, Muraru D, Badano LP, et al. Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain: The FAST-EFs Multicenter Study. J Am Coll Cardiol. 2015;66(13):1456-66.
34. Kim S, Park HB, Jeon J, Arsanjani R, Heo R, Lee SE, et al. Fully automated quantification of cardiac chamber and function assessment in 2-D echocardiography: clinical feasibility of deep learning-based algorithms. Int J Cardiovasc Imaging. 2022;38(5):1047-59.
35. Jiang J, Liu B, Li YW, Hothi SS. Clinical service evaluation of the feasibility and reproducibility of novel artificial intelligence based-echocardiographic quantification of global longitudinal strain and left ventricular ejection fraction in trastuzumab-treated patients. Front Cardiovasc Med. 2023;10:1250311.

36. Jang Y, Choi H, Yoon YE, Jeon J, Kim H, Kim J, et al. An Artificial Intelligence-Based Automated Echocardiographic Analysis: Enhancing Efficiency and Prognostic Evaluation in Patients With Revascularized STEMI. Korean Circ J. 2024;54(11):743-56.
37. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, et al. Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert. Circ Cardiovasc Imaging. 2019;12(9):e009303.
38. Sveric KM, Botan R, Dindane Z, Winkler A, Nowack T, Heitmann C, et al. Single-Site Experience with an Automated Artificial Intelligence Application for Left Ventricular Ejection Fraction Measurement in Echocardiography. Diagnostics. 2023;13(7):1298.
39. Li X, Liao L, Wu K, Meng AT, Jiang Y, Zhu Y, et al. An automatic and real-time echocardiography quality scoring system based on deep learning to improve reproducible assessment of left ventricular ejection fraction. Quant Imaging Med Surg. 2025;15(1):770-85.
40. Lin M, Zhang L, Wang Z, Liu H, Wang K, Tang G, et al. A combined system with convolutional neural networks and transformers for automated quantification of left ventricular ejection fraction from 2D echocardiographic images. Intelligent Medicine. 2025;5(1):46-53.
41. Lafitte S, Lafitte L, Jonveaux M, Pascual Z, Ternacle J, Dijos M, et al. Integrating artificial intelligence into an echocardiography department: Feasibility and comparative study of automated versus human measurements in a high-volume clinical setting. Archives of Cardiovascular Diseases. 2025.
42. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015;28(1):1-39.e14.
43. Glikson M, Nielsen JC, Kronborg MB, Michowitz Y, Auricchio A, Barbash IM, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy: Developed by the Task Force on cardiac pacing and cardiac resynchronization therapy of the European Society of Cardiology (ESC) With the special contribution of the European Heart Rhythm Association (EHRA). European Heart Journal. 2021;42(35):3427-520.
44. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation. 2018;138(16):1623-35.
45. Mor-Avi V, Khandheria B, Klempfner R, Cotella JI, Moreno M, Ignatowski D, et al. Real-Time Artificial Intelligence-Based Guidance of Echocardiographic Imaging by Novices: Image Quality and Suitability for Diagnostic Interpretation and Quantitative Analysis. Circ Cardiovasc Imaging. 2023;16(11):e015569.
46. Nagata Y, Kado Y, Onoue T, Otani K, Nakazono A, Otsuji Y, et al. Impact of image quality on reliability of the measurements of left ventricular systolic function and global longitudinal strain in 2D echocardiography. Echo Res Pract. 2018;5(1):27-39.
47. Ottenhoff J, Hewitt M, Makonnen N, Kongkatong M, Thom CD. Comparison of the Quality of Echocardiography Imaging Between the Left Lateral Decubitus and Supine Positions. Cureus. 2022;14(11):e31835.

48. Lyng Lindgren F, Tayal B, Bundgaard Ringgren K, Ascanius Jacobsen P, Hay Kragholm K, Zaremba T, et al. The variability of 2D and 3D transthoracic echocardiography applied in a general population : Intermodality, inter- and intraobserver variability. Int J Cardiovasc Imaging. 2022;38(10):2177-90.
49. Fotaki A, Puyol-Antón E, Chiribiri A, Botnar R, Pushparajah K, Prieto C. Artificial Intelligence in Cardiac MRI: Is Clinical Adoption Forthcoming? Front Cardiovasc Med. 2021;8:818765.
50. Tatsugami F, Nakaura T, Yanagawa M, Fujita S, Kamagata K, Ito R, et al. Recent advances in artificial intelligence for cardiac CT: Enhancing diagnosis and prognosis prediction. Diagnostic and Interventional Imaging. 2023;104(11):521-8.
51. McDonagh TAM, Marco; Adamo, Marianna; Gardner, Roy S.; Baumbach, Andreas; Böhm, Michael. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal - Cardiovascular Imaging. 2021;42(3599-3726).
52. Farsalinos KE, Daraban AM, Ünlü S, Thomas JD, Badano LP, Voigt J-U. Head-to-Head Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The EACVI/ASE Inter-Vendor Comparison Study. Journal of the American Society of Echocardiography. 2015;28(10):1171-81.e2.
53. Thavendiranathan P, Negishi T, Somerset E, Negishi K, Penicka M, Lemieux J, et al. Strain-Guided Management of Potentially Cardiotoxic Cancer Therapy. J Am Coll Cardiol. 2021;77(4):392-401.
54. Voigt JU, Pedrizzetti G, Lysyansky P, Marwick TH, Houle H, Baumann R, et al. Definitions for a common standard for 2D speckle tracking echocardiography: consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging. Eur Heart J Cardiovasc Imaging. 2015;16(1):1-11.
55. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, the Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making. 2020;20(1):310.
56. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again: Basic Books; 2019.
57. Balinisteanu A, Duchenne J, Puvrez A, Wouters L, Bézy S, Youssef A, et al. Vendor differences in 2D-speckle tracking global longitudinal strain: an update on a 10-year standardization effort. Eur Heart J Cardiovasc Imaging. 2025;26(8):1360-73.
58. Muraru D, Cecchetto A, Cucchini U, Zhou X, Lang RM, Romeo G, et al. Intervendor Consistency and Accuracy of Left Ventricular Volume Measurements Using Three-Dimensional Echocardiography. J Am Soc Echocardiogr. 2018;31(2):158-68.e1.
59. Strategic Market Research. Echocardiography Market Size, Global Analysis Report, 2030. 2022.
60. Yuda S, Sato Y, Abe K, Kawamukai M, Kouzu H, Muranaka A, et al. Inter-vendor variability of left ventricular volumes and strains determined by three-dimensional speckle tracking echocardiography. Echocardiography. 2014;31(5):597-604.
61. Chen T, Xia M, Huang Y, Jiao J, Wang Y. Cross-Domain Echocardiography Segmentation with Multi-Space Joint Adaptation. Sensors. 2023;23(3):1479.

62. Hernandez-Cruz N, Saha P, Sarker MMK, Noble JA. Review of Federated Learning and Machine Learning-Based Methods for Medical Image Analysis. Big Data and Cognitive Computing. 2024;8(9):99.
63. Baron T, Berglund L, Hedin E-M, Flachskampf FA. Test–retest reliability of new and conventional echocardiographic parameters of left ventricular systolic function. Clinical Research in Cardiology. 2019;108(4):355-65.
64. Sengupta PP, Dey D, Davies RH, Duchateau N, Yanamala N. Challenges for augmenting intelligence in cardiac imaging. Lancet Digit Health. 2024;6(10):e739-e48.
65. Fraser AG, Biasin E, Bijnens B, Bruining N, Caiani EG, Cobbaert K, et al. Artificial intelligence in medical device software and high-risk medical devices - a review of definitions, expert recommendations and regulatory initiatives. Expert Rev Med Devices. 2023;20(6):467-91.
66. Koçak B, Köse F, Keleş A, Şendur A, Meşe İ, Karagülle M. Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. Diagn Interv Radiol. 2025.
67. Morris DA. Clinical Relevance of Senior-Supervised Transthoracic Echocardiography in Clinical Practice and Research: An Editorial Commentary and Systematic Review. Echocardiography. 2025;42(1):e70085.
68. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60.
69. St Pierre SR, Peirlinck M, Kuhl E. Sex Matters: A Comprehensive Comparison of Female and Male Hearts. Front Physiol. 2022;13:831179.
70. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn (2020). 2020;2020:151-9.

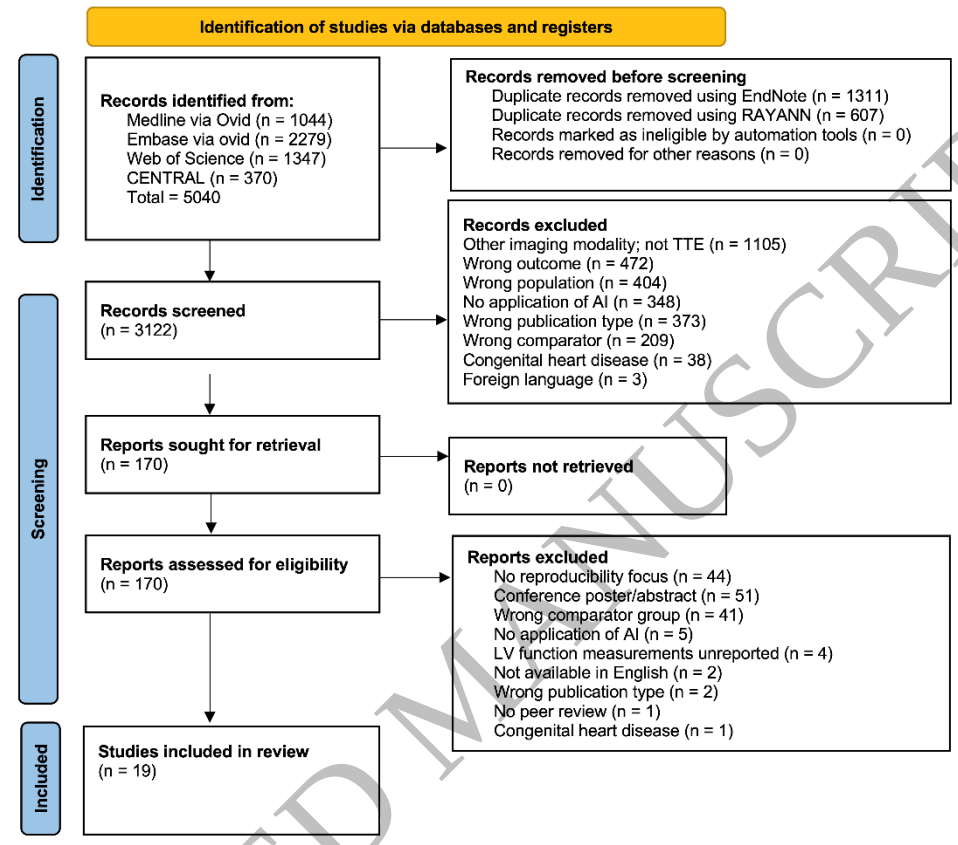
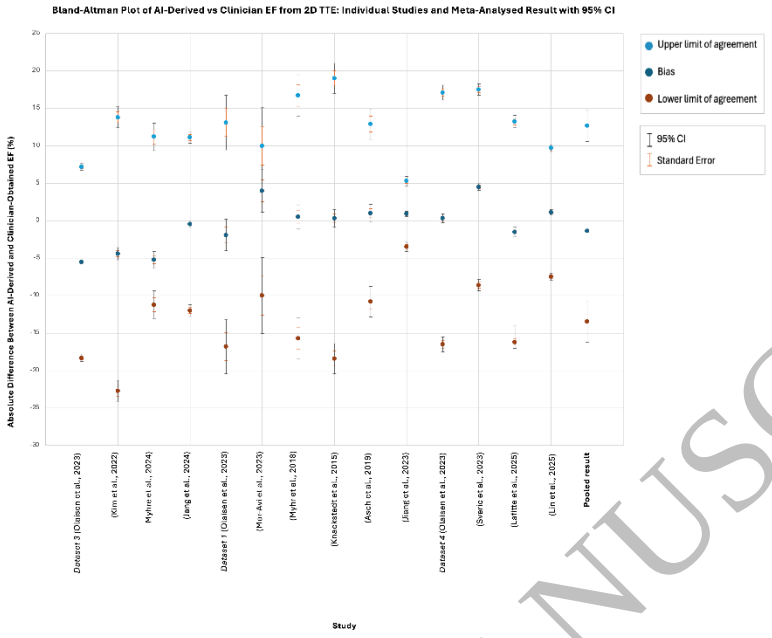


Figure 1
159x225 mm (x DPI)

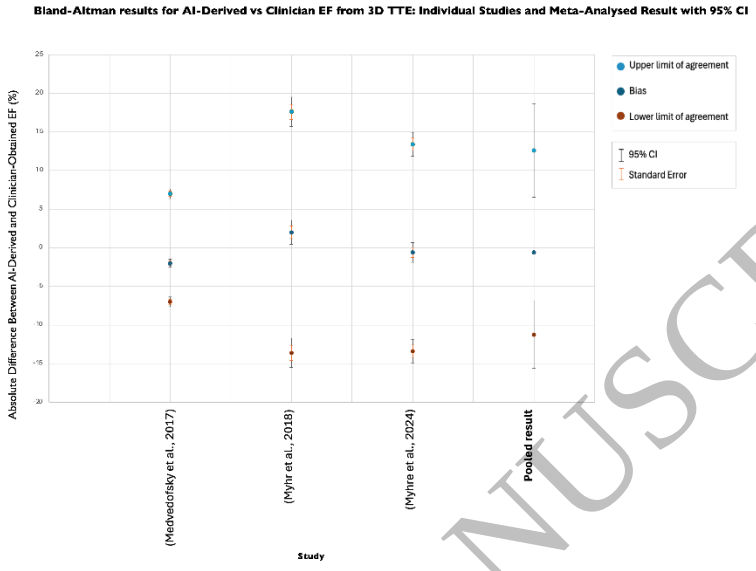


1

2 *Figure 2*

3 *159x113 mm (x DPI)*

4



1

2 *Figure 3*
3 *159x113 mm (x DPI)*

4

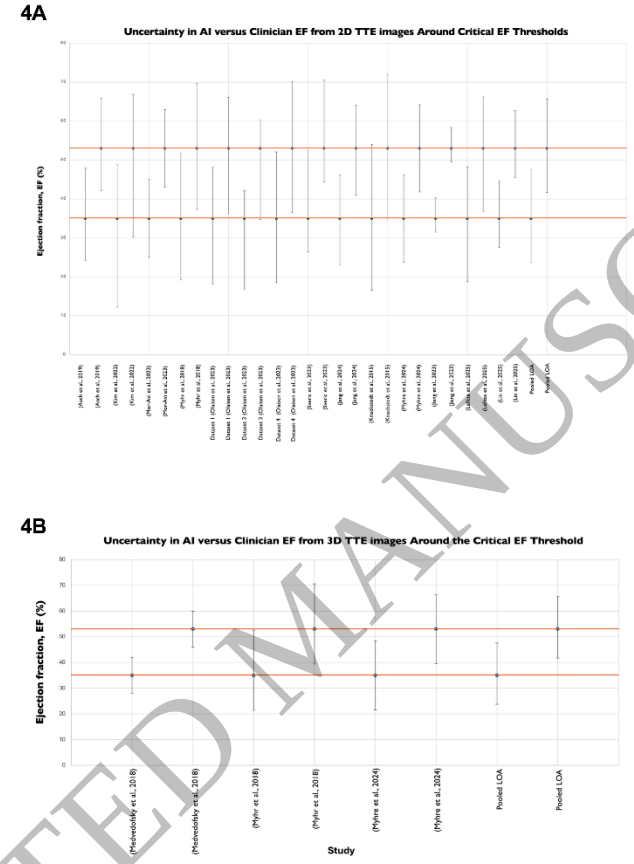


Figure 4
123x246 mm (x DPI)

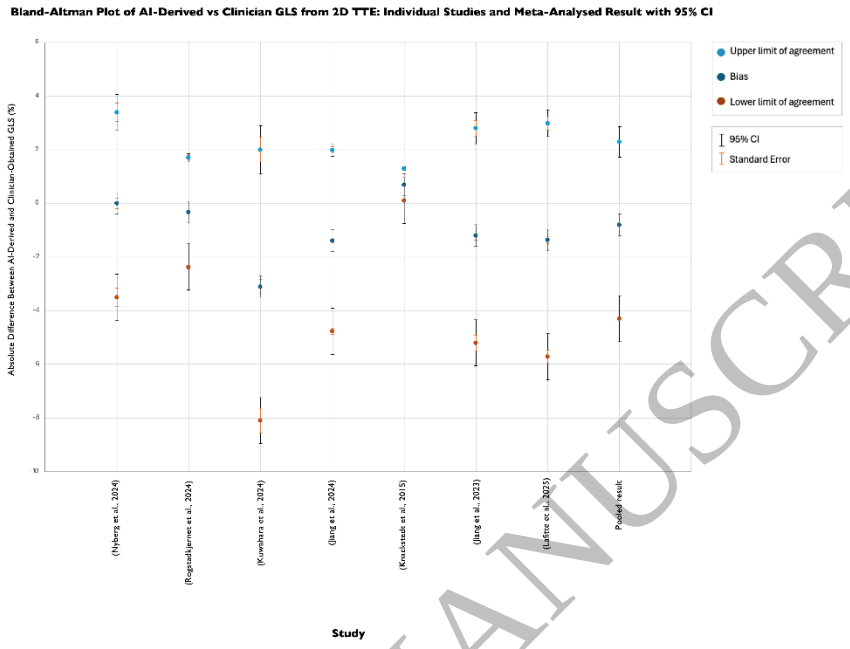


Figure 5
159x113 mm (x DPI)

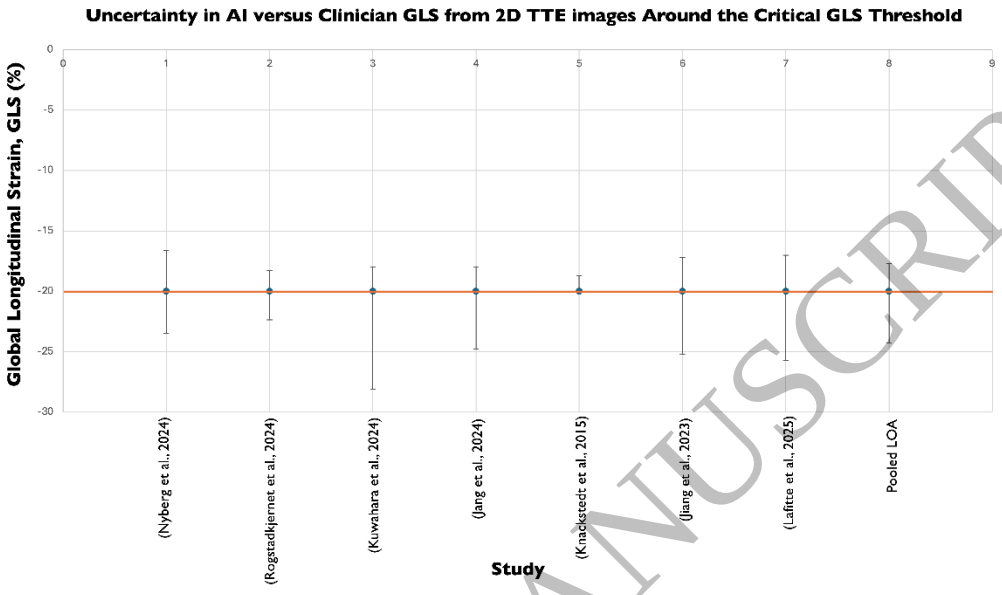
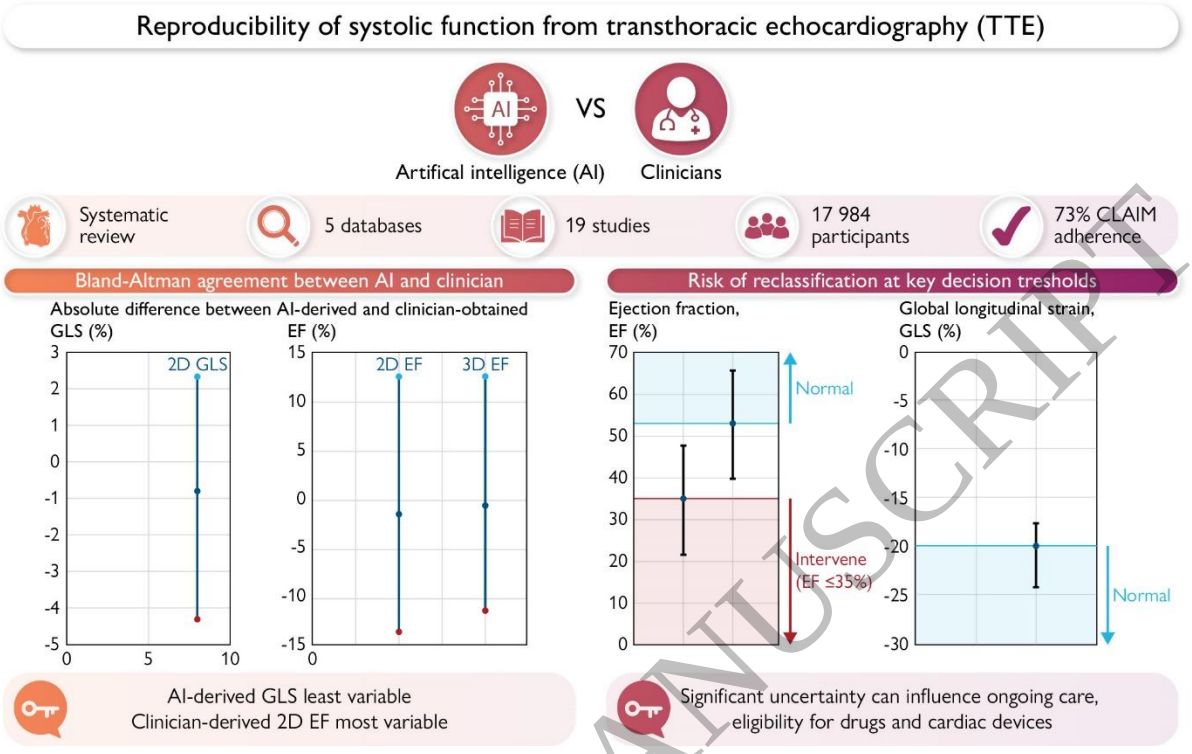


Figure 6
159x113 mm (x DPI)



Graphical Abstract