Review

DOI: 10.5582/ddt.2026.01020

Artificial intelligence (AI)-assisted ultrasound in clinical trials: Endpoint automation, decentralized monitoring, and regulatory readiness

Kenji Karako[1,*], Jianjun Gao[2]

[1] Department of Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan;
[2] Department of Pharmacology, School of Pharmacy, Qingdao Medical College, Qingdao University, Qingdao, Shandong, China.

SUMMARY: Ultrasound is among the most widely used imaging modalities in clinical trials, and yet its dependence on operator skill and equipment settings has historically limited the reproducibility of ultrasound-based endpoints in multi-center studies. Artificial intelligence (AI) now addresses this limitation across two complementary dimensions: automated measurement algorithms that quantify cardiac function, organ volume, and vascular parameters with reproducibility approaching or, in select settings, exceeding that of trained human readers and real-time acquisition guidance systems that enable clinicians with no formal sonography training to perform diagnostic-level examinations, making remote and decentralized assessment increasingly feasible. This narrative review synthesizes current evidence and regulatory developments across three interconnected domains. First, use of automated ultrasound to ascertain endpoints has advanced from single-institution validation to prospective and randomized evidence, with deep learning measurement of the left ventricular ejection fraction demonstrating formal equivalence to expert readers across multiple echocardiographic parameters and AI-first workflows shortening the time to diagnosis in a blinded non-inferiority trial. Second, AI-guided ultrasound acquisition by nurses and other non-expert operators has achieved a high rate of diagnostic acceptability in cardiac and pulmonary ultrasound, laying the groundwork for use of ultrasound-based endpoints in decentralized clinical trial designs, as reflected in Food and Drug Administration (FDA) guidance on decentralized trial elements. Third, the regulatory frameworks governing AI-enabled medical devices - including the U.S. FDA's Predetermined Change Control Plan guidance, the EU Artificial Intelligence Act, and internationally harmonized good machine learning practice relevant to Japan and other jurisdictions - increasingly emphasize overlapping principles such as specification of prospective performance, post-marketing oversight, and transparent reporting. Addressing remaining challenges in domain generalization across vendors, subgroup fairness, and algorithm change management during ongoing trials will be essential for AI-assisted ultrasound to fulfill its potential as a robust, scalable endpoint in clinical research worldwide.

Keywords: Artificial intelligence, ultrasound, clinical trials, decentralized clinical trials, automated endpoint measurement

1. Introduction

Ultrasound (US) imaging occupies a distinctive position as a clinical trial methodology. Unlike computed tomography or magnetic resonance imaging, US is non-ionizing, the equipment is portable, and scans can be acquired in real time, enabling its use across a spectrum of trial settings—from centralized imaging facilities in pivotal phase III studies to point-of-care assessments in resource-limited environments (*1*). Cardiac function endpoints such as the left ventricular ejection fraction (LVEF), as well as measurements of organ volume and vascular parameters, are routinely incorporated as primary or secondary endpoints in cardiovascular, oncological, and obstetric trials (*1*). The fact that clinical research is being performed around the world in settings with fewer resources, accelerated by disruptions to conventional trial infrastructure due to the COVID-19 pandemic, has further heightened the demand for imaging modalities that can function reliably outside tertiary hospitals.

Nonetheless, ultrasound has historically been the imaging modality most susceptible to measurement variability. Image quality and diagnostic accuracy depend heavily on operator skill, probe selection, patient positioning, and equipment settings—factors that vary

across sites, sonographers, and time points within the same trial. LVEF estimation using conventional two-dimensional echocardiography involves significant inter-observer variability. 2D methods result in minimal detectable differences of approximately 10–13%, a magnitude that may obscure meaningful treatment-related changes in cardiovascular outcome trials (*2*). Domain shift—the degradation of performance when a model is applied to data from a different institution, scanner, or patient population—is a further intrinsic challenge in ultrasound that constrains the external validity of both human readers and automated algorithms (*3,5*). These characteristics have limited the scalability of ultrasound-based endpoints in multi-center studies, where standardization of acquisition and interpretation is essential for valid cross-site comparisons.

Artificial intelligence (AI), and deep learning in particular, has emerged as a transformative tool to overcome these limitations. Convolutional neural networks (CNNs), vision transformers (ViTs), and self-supervised learning (SSL) frameworks have demonstrated the ability to automate ultrasound image analysis with reproducibility meeting or exceeding that of human experts when performing standardized tasks (*3,5,8,9*). A landmark video-based model for LVEF estimation had a mean absolute error (MAE) of 4.1% when using internal data and 6.0% when using an independent external dataset, with comparable inter-clinician variability, thus laying the foundation for automated echocardiography as an endpoint (*3*). A blinded, randomized non-inferiority trial that evaluated an AI-first echocardiographic workflow noted significantly lower rates of clinically meaningful LVEF discordance (> 5%) between initial assessment and the final cardiologist reading in the AI arm versus the sonographer arm (16.8% *vs.* 27.2%; $p < 0.001$ for superiority). This established a precedent for use of AI-assisted measurement as a primary trial endpoint in randomized controlled trials (RCTs) (*4*). In terms of ultrasound acquisition, AI-guided real-time probe-positioning systems have enabled clinicians with no prior sonography training—as well as nurses—to perform diagnostic-quality 10-view echocardiography, with 100% of primary-endpoint examinations meeting diagnostic acceptability criteria across all four primary endpoints in a prospective international noninferiority study (*6*). These advances in cardiac imaging suggest a shift in the central question from whether AI can assist ultrasound to how such assistance can be validated, standardized, and regulated within the formal framework of a clinical trial.

Three interrelated challenges define the current frontier. First, use of automated measurement algorithms to reliably ascertain clinical trial endpoints and whether they meet standards for accuracy, reproducibility, and external validity across heterogeneous scanning environments and patient populations encountered in multi-center studies must be validated. Second, AI-guided ultrasound acquisition foreshadows the potential for decentralized clinical trials (DCTs)—designs in which some or all imaging assessments are performed at local facilities or remotely rather than at designated imaging centers—but raises new questions about quality assurance, operator training, and data integrity. A regulatory framework for such designs was finalized by the U.S. Food and Drug Administration (FDA) in September 2024 (*7*). Third, the regulatory frameworks governing AI-based medical devices are evolving rapidly across major jurisdictions, and trial sponsors must navigate these requirements when incorporating AI-assisted ultrasound as an endpoint or trial tool.

This review addresses all three challenges systematically. We first examine the evidence base for using automated ultrasound in endpoint measurement, covering technical approaches, clinical validation studies up to and including RCTs, and the implications that a domain shift would have in multi-center deployment. We then consider AI-guided image acquisition in the context of decentralized trials and emergency setting trials, synthesizing prospective clinical evidence and identifying barriers to widespread implementation. Finally, we review the current regulatory landscape across the United States, Japan, and the European Union and outline the methodological standards—reporting guidelines, evaluation frameworks, and privacy-preserving learning architectures—that trial designers should incorporate to ensure reproducibility and regulatory compliance. Together, these three components represent the infrastructure required for AI-assisted ultrasound to fulfill its potential as a robust clinical trial tool. The conceptual relationship among these components is illustrated in Figure 1. The literature search strategy and inclusion criteria underlying this synthesis are described in the Methods section below.

2. Methods

This narrative review synthesizes recent literature on AI-assisted ultrasound in clinical trials. Relevant studies were identified through searches of PubMed, Embase, IEEE Xplore, and the Cochrane Library using combinations of the terms "ultrasound," "echocardiography," "artificial intelligence," "deep learning," "clinical trial," and "decentralized clinical trial." Additional sources were identified in references listed in key articles and from the websites of major regulatory agencies, including the FDA, Pharmaceuticals and Medical Devices Agency (PMDA)/Ministry of Health, Labor, and Welfare (MHLW), International Medical Device Regulators Forum (IMDRF), and the EU Official Journal.

Priority was given to prospective studies, randomized or non-inferiority trials, formal external validation studies, and primary regulatory guidance documents published up to March 2026. When preprints were cited,

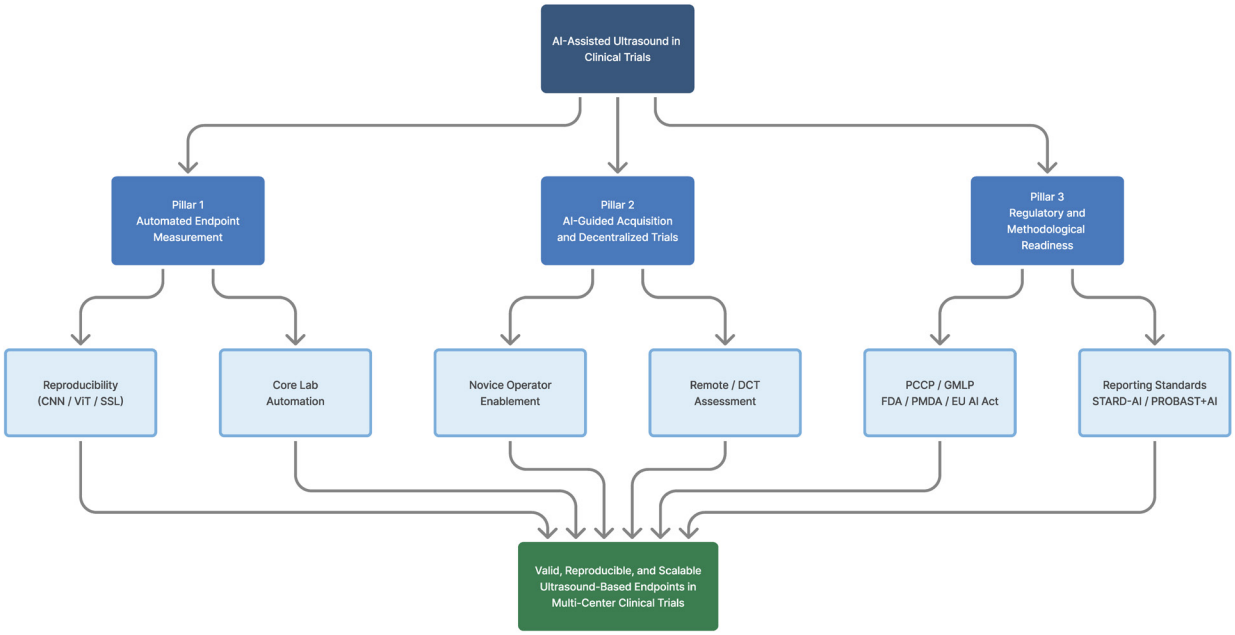

Figure 1. Conceptual framework of AI-assisted ultrasound in clinical trials. Three interdependent pillars are required for AI-assisted ultrasound to function as a reliable clinical trial tool. Pillar 1 (automated endpoint measurement) and Pillar 2 (AI-guided ultrasound acquisition and decentralized trials) each generate the evidence and infrastructure needed for valid endpoints. Pillar 3 (regulatory and methodological readiness) provides the governance layer that enables deployment across sites and regulatory acceptance. All three pillars converge on the goal of valid, reproducible, and scalable ultrasound-based endpoints in multi-center trials.

this was made explicit in the references and interpreted as emerging rather than definitive evidence. This was a narrative synthesis rather than a systematic review, so formal study selection, risk-of-bias adjudication, and meta-analytic pooling procedures were not performed.

3. Using automated ultrasound to ascertain clinical trial endpoints

3.1. Technological foundations

Three families of deep learning architecture underpin automated ultrasound measurement: CNNs, ViTs, and SSL frameworks. Each addresses a different constraint inherent to clinical ultrasound data.

CNNs were the first deep learning approach to ultrasound that was clinically validated. Their inductive biases—local feature detection and translational invariance—suit the spatially structured patterns of B-mode images, and their computational efficiency enables near-real-time inference on consumer hardware. The most consequential demonstration of CNN-based ultrasound analysis in a clinical trial context is EchoNet-Dynamic, which applied a video-based convolutional neural network to full echocardiographic video sequences to estimate the LVEF and segment the left ventricle (*3*). The model achieved an MAE of 4.1% on the internal held-out test set, which is well within the range of inter-observer variability among clinicians, and generalized to an independent external healthcare system with an MAE of 6.0%, demonstrating robust external validity for a video-based AI echocardiography system (*3*).

ViTs address a limitation of CNNs: their localized receptive fields hamper the integration of global spatial context, which is important for tasks such as view classification and whole-image quality assessment. By treating image patches as sequential tokens and applying self-attention mechanisms, ViTs capture long-range dependencies across the field of view. A systematic review of 69 publications covering ViT applications across cardiac, breast, prostate, and fetal ultrasound highlighted their strength in capturing global contextual information. At the same time, multiple studies reported that hybrid architectures combining convolutional feature extraction with transformer attention can improve performance, and particularly in settings with limited training data (*8*).

SSL exploits the large volumes of unlabeled ultrasound images available in clinical archives—data that are impractical to annotate exhaustively—by training models to reconstruct, contrast, or predict features of their own input without manual labels. This approach is especially valuable for reducing dependence on expert annotation in specialized domains. At the foundation model scale, UltraFedFM pre-trained a shared model *via* federated learning across 16 institutions in 9 countries using over one million unlabeled ultrasound images, achieving an area under the receiver operating characteristic curve (AUROC) of 0.927 for disease classification and a Dice coefficient of 0.878 for lesion segmentation across downstream tasks (*9*). The implications for multi-center trial infrastructure

are discussed further in Section 5. The principal characteristics, ultrasound-specific limitations, and representative evidence for each of the three families of architecture are compared in Table 1.

Cutting across all three families of architecture is the emerging field of explainable AI (XAI), which generates human-interpretable rationales for model predictions—for example, attention heatmaps highlighting the myocardial regions driving an LVEF estimate, or saliency maps indicating which image zones triggered a lesion classification. XAI is not an alternative architecture but a post-hoc or intrinsic interpretability layer that can be applied to CNN, ViT, and SSL-derived models alike. Its clinical trial relevance is twofold: regulators increasingly expect AI-enabled medical devices to provide interpretable outputs sufficient for clinical oversight, and FUTURE-AI's explainability dimension recommends that the basis for an AI decision be communicated to the clinical reviewer in a way that supports, rather than replaces, human judgment (*17*). For trial endpoints specifically, XAI outputs can serve as audit evidence during regulatory review of discordant AI-human measurements and as a mechanism for detecting systematic model failure modes—such as dependence on image artifacts rather than genuine anatomical features—before adoption.

3.2. Clinical validation of using AI to ascertain trial endpoints

Converting an AI measurement algorithm from conjecture into a validated clinical trial tool requires a structured hierarchy of evidence. This section first presents a pragmatic five-level framework synthesized from the literature on AI-assisted echocardiography—where the evidence base is currently most complete—and then considers how far analogous evidence has emerged in non-cardiac areas. Representative studies are mapped to each level in Figure 2.

Level 1—Internal validation using held-out data. The minimum requirement for any AI measurement system is that its performance has been determined using a held-out test set from the same institution or dataset as training data. When estimating the LVEF, EchoNet-Dynamic achieved an MAE of 4.1% with its internal held-out set—approaching the inter-clinician variability for the LVEF reported in the same clinical environment—establishing a minimum credibility threshold (*3*).

Level 2—External validity using independent data. In addition, EchoNet-Dynamic's performance in an independent external healthcare system with different scanning environments yielded an MAE of 6.0%—

Table 1. The three families of architecture in terms of typical endpoints, key limitations in the ultrasound context, and representative references

Architecture	Typical endpoints	Ultrasound-specific limitations	Representative reference
CNN (2D/3D)	LVEF, segmentation, view classification	Device–vendor domain shift; limited global context	Ouyang *et al*., Nature 2020 (*3*)
Vision Transformer	Classification, segmentation, quality grading	Requires larger training sets; sensitive to domain shift	Vafaeezadeh *et al*., Diagnostics 2024 (*8*)
SSL/Foundation model	Multi-task pre-training → downstream fine-tuning	Pre-training dataset diversity is a bottleneck	Jiang *et al*., NPJ Digit Med 2025 (*9*)

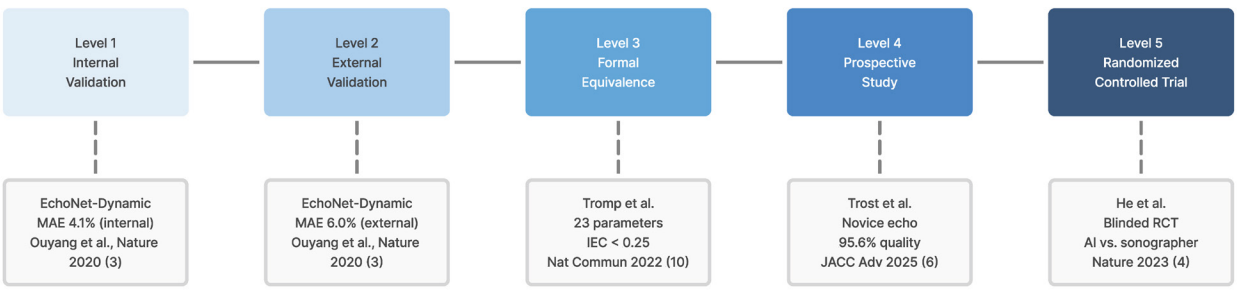

Figure 2. Hierarchy of clinical evidence for use of AI-assisted ultrasound to ascertain endpoints. Five levels of evidence for use of AI-assisted ultrasound to ascertain clinical trial endpoints. They are shown here, along with landmark studies in AI-assisted echocardiography. Level 1 involves establishing internal performance based on held-out data from the institution developing the AI; Level 2 involves demonstrating external validity using independent data from a separate institution or scanning environment; Level 3 involves applying a pre-specified equivalence criterion (*e.g.*, the individual equivalence coefficient) to repeated human expert assessments; Level 4 involves prospectively evaluating ultrasound performed by non-expert operators under trial conditions; and Level 5 indicates the steps in a randomized, blinded evaluation. A trial sponsor should pre-specify the minimum level of evidence required for endpoint adoption in the trial protocol; Level 3 is the recommended minimum for primary endpoints.

within the range of human inter-clinician variability (*3*). This constitutes the first large-scale demonstration of cross-site external validity for a video-based AI model for echocardiography and establishes the principle that cross-site validation should be prospectively planned rather than retrospective. An algorithm remains at Level 1 until this external step has been completed.

Level 3—Formal equivalence validation. Tromp *et al.* applied a more rigorous standard, comparing deep learning measurements of 23 echocardiographic parameters—including the ejection fraction, ventricular volume, and Doppler-derived filling pressure—from 600 participants to three independent repeated measurements by trained sonographers (*10*). The primary criterion was the individual equivalence coefficient, a statistic that quantifies whether algorithm-human disagreement is smaller than human-human disagreement. For all 23 parameters, the upper bound of the 95% confidence interval (CI) for the individual equivalence coefficient fell below the pre-specified non-inferiority threshold of 0.25, indicating that the deep learning workflow was at least as reproducible as expert human assessment (*10*). This study design, which pre-registers the equivalence criterion before data analysis, provides the most relevant methodological template for regulatory and endpoint validation purposes.

Level 4—Prospective evaluation in the intended operator population. An algorithm validated at Level 3 has been assessed in a controlled laboratory setting but not necessarily in the hands of the operator population that will use it in a trial. Trost *et al.* addressed this gap in a two-site international study (240 participants) in which nurses—the intended DCT operator—used an AI-guided system to perform echocardiography, achieving 100% diagnostic acceptability for all four pre-specified primary endpoints (LV size, global LV function, RV size, and pericardial effusion), identical to the expert reference group (*6*). This level of evidence is analogous to a phase II clinical study for a drug: it confirms that a validated algorithm remains fit-for-purpose when used by non-expert operators under real trial conditions. That study is discussed in detail in Section 4.2.

Level 5—Randomized controlled trial of the AI-enhanced workflow. He *et al.* conducted a blinded, randomized non-inferiority trial at Cedars-Sinai Medical Center comparing AI-assisted initial LVEF assessment to standard sonographer assessment (*4*). The pre-specified primary endpoint was the proportion of studies in which the initial assessment (AI or sonographer) differed from the reviewing cardiologist's final LVEF reading by more than 5%. Among 3,495 studies, this discordance rate was significantly lower in the AI arm (16.8%) than in the sonographer arm (27.2%; difference −10.4%, 95% CI −13.2 to −7.7; $p < 0.001$ for superiority), and blinding was successfully maintained (blinding index: 0.088) (*4*). This trial demonstrated that an AI-first echocardiographic workflow can be evaluated within the formal infrastructure of a randomized, blinded clinical study—a critical precedent for trials seeking to use AI measurement as their primary endpoint.

3.2.1. Generalizability to non-cardiac ultrasound domains

The five-level hierarchy described above is most clearly evident in AI-assisted echocardiography, where the sequence from internal validation to RCT evidence now spans multiple independent groups. Evidence from non-cardiac domains is accumulating, though most applications remain at Levels 1–3. A prospective two-country diagnostic accuracy study of fetal ultrasound demonstrated that deep learning estimation of gestational age from blind ultrasound sweeps achieved an MAE of 3.2 days. The AI system met a pre-specified equivalence margin (± 2 days) relative to standard sonographer-performed biometry when both were evaluated with first-trimester crown–rump length as the reference standard (*26*). A complementary four-center prospective study conducted across Moroccan healthcare centers developed and prospectively evaluated a deep learning system for end-to-end automation of fetal biometry and amniotic fluid assessment from ultrasound cine-loops. The model achieved limits of agreement with expert measurements that were narrower than reported intra- and inter-observer variability among sonographers, indicating reproducibility comparable to or better than human performance (*27*). A six-center prospective multicenter study of hepatic ultrasound developed and internally validated a deep learning model for multi-stage liver fibrosis classification (S0–S4) using paired high-frequency ultrasound images of the liver and spleen, with histologic staging as the reference standard. The model achieved an AUROC of 0.87–0.94 across clinically relevant fibrosis thresholds and demonstrated the potential to reduce biopsy rates in simulated clinical pathways (*28*). The Controlled ARTHUR Trial of musculoskeletal ultrasound prospectively evaluated a robotic ultrasonography system combined with AI-based synovitis scoring in rheumatoid arthritis, comparing its performance to that of expert-performed ultrasound. The study noted moderate agreement between robotic and human assessments (intraclass correlation coefficients of 0.59–0.64) and limited diagnostic accuracy (approximately 59%), highlighting both the feasibility and current limitations of operator-independent ultrasound acquisition and automated scoring (*29*). Across these domains, emerging evidence suggests that AI-based measurement can achieve levels of reproducibility comparable to expert human assessment in selected applications, while prospective multi-center studies are beginning to establish the Level 2–3 evidence base required for further validation for clinical and trial use.

3.3. Challenges in multi-center deployment

Despite this evidence, three challenges constrain the use of automated ultrasound to ascertain endpoints across heterogeneous multi-center trial networks.

Domain shift—performance degradation due to differences in the scanner vendor, probe frequency, imaging protocol, or patient population—is the most consistently cited obstacle. Recently, Universal UltraSound Image Challenge 2025 evaluated general-purpose ultrasound AI models' performance on data aggregated from public datasets and partner hospitals across multiple anatomical regions, and the competition revealed measurable degradation of performance on data from institutions not represented in the training dataset (*11*). Although this evidence is still preliminary, it reinforces an important operational point: domain generalization must be explicitly validated, not assumed, before multi-center deployment.

Subgroup performance gaps are a related concern. Differences in AI measurement accuracy across patient subgroups—potentially defined by sex, age, body mass index, or scanner type—may introduce systematic bias that inflates or attenuates between-arm differences in a trial with imbalanced enrollment, analogous to differential misclassification in conventional measurement. Sex-related performance gaps have been empirically documented in musculoskeletal ultrasound AI (*12*); analogous biases are plausible across cardiac, hepatic, and vascular domains, where training dataset composition is similarly skewed. Pre-registration of analyses of subgroup accuracy and reporting of subgroup-stratified performance metrics are therefore essential components of validating use of AI to ascertain endpoints.

Public data availability remains a bottleneck for algorithm development and replication. A systematic catalog of publicly available ultrasound resources identified 72 datasets and 56 open-source models as of September 2025, with notable underrepresentation of fetal (5 datasets) and prostate (4 datasets) content relative to the frequency of their clinical use (*13*). For investigators seeking to independently validate algorithms before adoption in a trial, this scarcity of benchmark data is a practical impediment.

3.4. Reporting standards

Transparent reporting is a prerequisite for regulators, site monitors, and independent reviewers to evaluate whether an AI endpoint meets trial quality standards. Two complementary guidelines are particularly relevant. STARD-AI, published in Nature Medicine in 2025, extends the established STARD checklist with AI-specific reporting items covering dataset composition, model evaluation, and bias or fairness considerations (*14*). PROBAST+AI, published in the British Medical Journal in 2025, provides a structured tool for assessing risk of bias and applicability in predictive model studies using regression or AI (*15*). These frameworks do not themselves constitute regulatory requirements, but they increasingly shape the evidence packages that regulators, sponsors, and reviewers expect when proposing use of an AI system in a trial. The alignment between common trial study designs and their most relevant reporting or evaluation framework is outlined in Table 2.

4. AI-guided ultrasound acquisition and decentralized ultrasound monitoring

4.1. Mechanisms of AI-guided ultrasound acquisition and their significance for clinical trials

Automated measurement algorithms, however accurate, are only as reliable as the images on which they operate. Suboptimal image acquisition—incorrect probe orientation, inadequate acoustic coupling, or failure to obtain the prescribed view—introduces a source of variability that no downstream algorithm can fully correct. AI-guided ultrasound addresses this upstream constraint by providing real-time, computer-generated feedback to the operator during image acquisition, decoupling diagnostic quality from individual sonographer experience.

The technical architecture of guided image acquisition systems typically combines two deep learning components: a view classification network that identifies which of the target views has been achieved, and a guidance-providing module that translates the current probe position into actionable instructions—expressed as directional cues on a display or auditory prompts—to direct the operator toward the correct plane. These components operate on live video input with inference latencies low enough for interactive use in clinical settings in real time. The classification network may be a CNN, a ViT, or a hybrid architecture; the critical design requirement is the ability to distinguish near-correct views from target views with sufficient specificity to avoid issuing conflicting instructions, which is particularly challenging in echocardiography given the anatomical proximity of standard imaging planes.

Table 2. Trial study designs within the most relevant reporting or evaluation frameworks

Study type	Primary framework	Key AI-specific requirements
Diagnostic accuracy study	STARD-AI (*14*)	Dataset provenance, subgroup reporting, bias/fairness
Predictive model development/validation	PROBAST+AI (*15*)	Risk of bias across participants, predictors, outcomes, analysis
Early-stage real-world evaluation	DECIDE-AI (*16*)	Human–AI interaction, safety monitoring
Comprehensive evaluation of implementation	FUTURE-AI (*17*)	Fairness, universality, traceability, robustness, explainability

When conducting a clinical trial, the significance of reliable AI-guided ultrasound acquisition extends beyond quality control into the trial design itself. If novice operators—nurses, first-responders, or community health workers without formal sonography training—can obtain diagnostic-quality ultrasound images with AI guidance, then imaging assessments that would previously have required a patient to travel to an imaging facility can instead be performed at local clinics, remote research sites, or in participants' homes. This capability is a prerequisite for a DCT involving imaging-dependent endpoints and directly addresses the operational constraints that have historically limited ultrasound-based endpoints to site-intensive trial designs.

4.2. Prospective clinical evidence

The clinical evidence base for AI-guided ultrasound acquisition has expanded substantially in the past two years and now spans cardiac, pulmonary, vascular, obstetric, and abdominal applications.

Cardiac ultrasound. The most rigorous assessment to date is a prospective, international, noninferiority study conducted at two sites (United States and France) by Trost *et al*. (*6*). Nurses without prior echocardiography experience used an AI-guided system to acquire 10 standard transthoracic echocardiographic views of 240 patients over a nine-month period (November 2023 to August 2024). For all four pre-specified primary endpoints—assessment of left ventricular size, global left ventricular function, right ventricular size, and nontrivial pericardial effusion—100% of novice-performed examinations met diagnostic acceptability criteria, with no difference compared to expert-performed examinations. Examinations were independently reviewed by five cardiologists blinded to operator identity, and strong agreement was observed between AI-guided image acquisition by novices and that by experts across multiple parameters. Of methodological note, the study prospectively defined its performance thresholds and employed blinded expert adjudication, providing high-quality prospective evidence supporting the feasibility of AI-guided ultrasound by non-expert operators in clinical settings.

Pulmonary ultrasound. Baloescu *et al*. extended the acquisition-guidance paradigm to pulmonary ultrasound in a multicenter study spanning four sites (*18*). Non-expert operators used the system to perform a standardized eight-zone protocol in patients with acute respiratory symptoms. Diagnostic-quality images were obtained in 98.3% of examinations, with no significant difference compared to expert-performed examinations. Importantly, the study focused on acquisition quality rather than diagnostic accuracy, demonstrating that AI guidance enables non-expert operators to reliably obtain images suitable for clinical interpretation. These findings are relevant to decentralized trial settings, where imaging must be performed by a distributed workforce without extensive ultrasound expertise.

Vascular ultrasound. Speranza *et al.* evaluated AI guidance in diagnosis of deep vein thrombosis in a setting where non-expert operators performed compression ultrasound of the lower extremities (*19*). In a multicenter study of 381 patients across 11 UK hospitals, scans acquired with AI guidance were reviewed by qualified clinicians and compared to standard examinations by imaging specialists. Emergency medicine reviewers achieved a sensitivity of 95-98% and a specificity of 97-100%. Importantly, the diagnostic workflow incorporated clinician review, with indeterminate or insufficient-quality scans requiring further assessment. This study highlights that AI-guided ultrasound can be integrated with expert review to maintain diagnostic performance while potentially reducing the need for universal expert-performed examinations.

Obstetric ultrasound. The acquisition-guidance paradigm extends to obstetric ultrasound, where access to trained sonographers is a primary constraint on the feasibility of maternal-fetal medicine trials in limited-resource settings. The prospective two-center study by Stringer *et al.* described in Section 3.2.1—in which novice operators with minimal training in Zambia performed freehand blind sweeps using an AI-guided portable device—yielded gestational age estimation that was non-inferior to that of credentialed sonographers, confirming that AI guidance can substitute for formal training with an obstetric endpoint that is image acquisition-dependent (*26*). Gomes *et al.* validated a complementary mobile-optimized AI system capable of running without Internet connectivity on a handheld device, designed specifically for novice operator use in resource-limited settings (*31*). The system provided feedback scores reflecting sweep quality, which can assist operators in identifying inadequate acquisitions, and achieved gestational age estimation to within -1.4 ± 4.5 days of standard sonographer biometry; fetal malpresentation—a clinically important finding that would qualify as a secondary endpoint in preterm birth prevention trials—was detected with an AUROC of 0.977 (95% CI 0.949–1.000) (*31*). Together these studies demonstrate that operator-independent obstetric imaging protocols are feasible for endpoints central to global reproductive health trials, including standardization of gestational age and screening for malpresentation.

Abdominal ultrasound. Chiu *et al*. conducted a prospective evaluation of AI-guided abdominal aortic aneurysm screening in which 10 nurses with no prior ultrasonography experience each performed fifteen scans using a real-time deep learning detection algorithm (*32*). Diagnostic image quality was achieved in 87.5% of nurse-performed scans, compared to 91.3% for physician-performed scans ($p = 0.310$), establishing equivalence at a clinically acceptable threshold. Abdominal aortic aneurysm detection sensitivity was

100% and specificity was 97.8% (AUROC 0.975), with an MAE for measurement of the aortic diameter of 2.8 mm—within the threshold conventionally used for surveillance decisions in screening programs. These findings suggest that AI-guided ultrasound acquisition may enable non-expert operators to perform abdominal ultrasound screening with performance comparable to experienced physicians, supporting its potential applicability in scalable screening and decentralized trial settings.

4.3. DCTs: Regulatory framework and implementation requirements

The FDA's September 2024 final guidance on conducting clinical trials with decentralized elements defines DCTs as studies in which trial-related activities occur outside traditional clinical trial sites, including through local healthcare providers, in-home visits, or telehealth and digital health technologies (*7*). The guidance recognizes that imaging and other clinical assessments may be performed at local facilities and emphasizes the need for standardized protocols, appropriate training of personnel, and robust data quality and integrity monitoring. Sponsors are also expected to specify data origin, data flow, and monitoring procedures within the trial protocol.

For ultrasound-based imaging endpoints specifically, the FDA's earlier guidance on Clinical Trial Imaging Endpoint Process Standards, finalized in 2018, outlines recommended process standards applicable across clinical trial designs (*20*). These include pre-specified imaging acquisition protocols, operator qualification and training, equipment standardization and quality control procedures, as well as requirements for image transfer, storage, and centralized interpretation where appropriate (*20*). Sponsors incorporating AI-guided ultrasound acquisition into decentralized trial designs should ensure that image quality and protocol adherence are maintained when performed by the intended user population.

Despite regulatory clarity and mounting clinical evidence, substantial barriers to implementation remain. The COMPASS-AI survey, which collected structured responses from 1,154 healthcare professionals across multiple countries, found that 81.1% expressed enthusiasm for AI-assisted POCUS, and yet identified training and education (27.1% of respondents) and clinical validation and evidence (17.5%) as the two most frequent barriers to adoption (*21*). Additional concerns included limitations in infrastructure, workflow integration, lack of standardized guidelines, and uncertainty regarding liability. These findings suggest that, in DCT settings, institutional readiness may represent a critical limiting factor.

4.4. Emergency and resource-constrained settings

The evidence reviewed in Sections 4.2 and 4.3 addresses the operational challenges of decentralizing clinical trials within established healthcare systems; the same AI-guided ultrasound acquisition capabilities also extend the scope of the evidence reviewed in Section 4.2 into emergency and resource-constrained settings, where conventional sonographer-dependent protocols are entirely infeasible and where endpoint integrity would otherwise be unachievable.

The COVID-19 pandemic provided a real-world stress test of decentralized imaging workflows. Widespread adoption of POCUS for cardiopulmonary triage in emergency departments and temporary care settings—where radiological infrastructure was limited and trained sonographers were scarce—demonstrated that portable ultrasound can be used effectively with simplified, protocolized acquisition approaches under emergency conditions. Although primarily used for clinical decision-making rather than formal trial endpoints, these experiences suggest that structured and protocolized ultrasound workflows can be implemented with practical feasibility in resource-constrained environments. This has important implications for decentralized trial design: AI-guided ultrasound acquisition systems may further enhance standardization and reproducibility, particularly when operator expertise and participant mobility are limited.

In low- and middle-income countries (LMICs), endpoint validity carries an additional dimension: the underrepresentation of LMIC populations in clinical trials is a recognized threat to the external validity and generalizability of trial results. A scoping review of 29 studies on AI-enabled POCUS applications in limited-resource settings identified pulmonology medicine and obstetrics as the most frequently investigated domains, with many studies focusing on diagnostic support and workflow facilitation (*22*). Implementation constraints in LMICs differ from those in high-income countries, with challenges including infrastructure limitations, lack of standardization, and limited data availability rather than solely operator training. The fetal ultrasound validation data from Zambia (*26*) and Morocco (*27*) reviewed in Section 3.2.1 provide direct evidence that AI-enabled systems can achieve performance comparable to expert standards in limited-resource environments. While the Zambian study demonstrated the equivalence of AI-assisted image acquisition by novice operators, the Moroccan study showed that automated measurement pipelines can reach human-level reproducibility. Together, those studies support the feasibility of scalable trial endpoints in LMICs.

5. Regulatory and methodological readiness

5.1. Evolving regulatory frameworks for AI-enabled medical devices

AI-based medical devices, including those with

automated ultrasound or guided image acquisition, are subject to regulatory oversight in all major jurisdictions. The regulatory treatment of these devices is undergoing rapid change as authorities develop frameworks that accommodate the unique characteristics of AI: continuous learning from post-deployment data, performance that may shift over time, and decision processes that resist traditional static software validation. An overview of the key regulatory features across all three jurisdictions is provided in Table 3.

United States. The FDA regulates AI-enabled ultrasound devices as Software as a Medical Device (SaMD) under the device provisions of the Federal Food, Drug, and Cosmetic Act. As of March 2026, the FDA had authorized more than 1,400 AI-enabled medical devices (*33*). The central regulatory challenge for AI-assisted endpoints is managing modifications: unlike conventional software, AI/ML models may require iterative updates as new training data become available, and each modification that could affect safety or effectiveness may trigger additional regulatory review under traditional 510(k) or De Novo pathways. The Predetermined Change Control Plan (PCCP) framework, for which the FDA finalized guidance in December 2024, addresses this challenge by allowing sponsors to pre-specify the categories of changes they anticipate making to an algorithm—such as retraining on additional data or refining post-processing thresholds—and to obtain prospective FDA agreement that such changes may be implemented without a new marketing submission, provided performance monitoring confirms they remain within pre-specified bounds (*23*). The Good Machine Learning Practice (GMLP) guiding principles, first articulated jointly by the FDA, Health Canada, and the UK's Medicines and Healthcare products Regulatory Agency and later elaborated through IMDRF, provide upstream development expectations applicable across AI/ML medical devices regardless of pathway (*25,30*). These principles cover data management, reference standard selection, and specification of clinically meaningful performance targets, and they complement the PCCP by defining the quality baseline against which predetermined changes should be evaluated.

Japan. Japan's regulatory approach to AI-enabled medical devices has developed alongside international SaMD and GMLP frameworks while retaining its own approval and consultation pathways. The original GMLP guiding principles were published jointly by the U.S. FDA, Health Canada, and the UK's Medicines and Healthcare products Regulatory Agency in 2021 (*25*) and were subsequently reflected in IMDRF guidance, in whose activities Japan participates (*30*). In Japan, AI-enabled medical devices are reviewed within the existing medical device regulatory framework administered by the MHLW and PMDA (*34*), and PMDA has established dedicated review and consultation pathways for SaMD (*35*). Risk management considerations are informed by internationally recognized medical device standards, including ISO 14971, alongside SaMD concepts developed through IMDRF. For qualifying innovative products, the SAKIGAKE Designation System may offer priority consultation and review for devices addressing unmet clinical needs (*36*).

European Union. The EU Artificial Intelligence Act (Regulation (EU) 2024/1689), which entered into force on August 1, 2024, applies on a phased timetable (*24*). The Act becomes generally applicable on August 2, 2026, while high-risk AI systems embedded in regulated products such as medical devices have an extended transition period until August 2, 2027 (*24*). AI-enabled ultrasound systems that guide image acquisition or automate measurement will usually need to satisfy the AI Act along with the Medical Device Regulation (MDR) or In Vitro Diagnostic Regulation (IVDR), rather than instead of them. This means sponsors and developers should anticipate requirements for technical documentation, risk management, human oversight, and post-market monitoring that are complementary to CE-marking obligations under product law. The phased implementation timeline gives developers a defined window to adapt existing systems, but it also means that regulatory planning for EU-inclusive trials should begin well before the 2027 deadline.

5.2. Implications for clinical trial design

The regulatory frameworks described above have direct implications for how AI-assisted ultrasound will be used to ascertain endpoints throughout the duration of a clinical trial.

PCCP and intra-trial model changes. A critical but underappreciated issue in trials using AI to ascertain

Table 3. The key regulatory features across the three jurisdictions

Jurisdiction	Key framework	Change management	Data requirements
United States (FDA)	SaMD / PCCP (*23*)	Predetermined changes pre-agreed with the FDA	GMLP principles (*25*)
Japan (PMDA)	SaMD / ISO 14971 (*34,35*)	Variation submission; IMDRF GMLP alignment (*30*)	Pseudonymized data guidelines (MHLW)
European Union	EU AI Act (*24*) + MDR/IVDR	Post-market monitoring + PCCP equivalent	Technical documentation per AI Act Annex IV

endpoints is algorithm stability. If a trial spans two or more years—as is typical for cardiovascular outcome studies—the sponsor may wish to retrain or update the algorithm on expanded data during the trial period to correct drift or improve accuracy in newly enrolled subpopulations. Under conventional validation paradigms, any modification would require re-validation of the endpoint definition and potentially revision of the statistical analysis plan. The PCCP framework provides a mechanism for managing such changes in a pre-specified, regulatory-compliant manner: if the scope and performance bounds of anticipated modifications are agreed upon with the regulator before trial initiation, updates can proceed without undermining endpoint integrity (*23*). Trial sponsors should therefore consider developing a PCCP for endpoints ascertained using AI in parallel with protocol development, treating algorithm change management as an essential element of defining endpoints.

Post-marketing performance monitoring. FDA requirements and the emerging EU AI Act compliance framework both point toward ongoing performance monitoring of AI systems while in use. For trial sponsors using AI-assisted endpoints, this translates into continued collection of performance data—often through a random or risk-based sample of AI measurements compared against a human reference—throughout the study. Integrating this monitoring into the workflow of the central imaging laboratory adds operational complexity but can provide early warning if systematic performance degradation is detected.

Statistical analysis plan (SAP) considerations for AI endpoints. A frequently overlooked dimension of incorporating AI measurements into clinical trials is their pre-specification in the SAP. Unlike conventional imaging endpoints, AI measurement algorithms produce point estimates that may be accompanied by model confidence scores or prediction intervals; the SAP must specify whether these uncertainty outputs will be used as covariates, to exclude low-confidence measurements from the primary analysis, or to trigger adjudication by a human reader. The equivalence or non-inferiority margin for the AI endpoint should be determined in consultation with regulators before database lock and should be stated explicitly in the SAP rather than inferred post-hoc. When the AI measurement serves as a surrogate endpoint, the statistical linkage between the surrogate and the clinical outcome of interest should be defined and cited in the protocol, following the Prentice criteria or an equivalent framework for surrogate endpoint validation. Finally, subgroup performance analyses—stratified by sex, age, body mass index, and scanner type—should be pre-specified as secondary endpoints rather than exploratory analyses, given documented evidence of differential AI performance in musculoskeletal ultrasound (*12*) and the recognized risk of analogous biases across imaging domains.

5.3. Federated learning as a regulatory-compliant infrastructure

The four reporting and evaluation frameworks described in Section 3.4—STARD-AI, PROBAST+AI, DECIDE-AI, and FUTURE-AI—have important implications for regulatory preparedness even though they are not, by themselves, binding regulatory instruments. Applied prospectively rather than retrospectively, they help shape the pre-approval evidence package and the post-deployment monitoring structure that regulators and sponsors increasingly expect. FUTURE-AI's emphasis on traceability is especially relevant: algorithm version, training data provenance, and inference settings should be logged for endpoint-generating systems so that discrepancies can be audited if they arise during trial oversight or regulatory review.

Federated learning architectures provide a technical mechanism through which PCCP-defined model updates may be implemented across sites, rather than directly fulfilling the PCCP requirements themselves. By enabling each trial site to compute local model updates on its own patient data and contribute only aggregated gradient information to a central server, federated learning allows iterative model refinement without centralizing identifiable patient records, substantially reducing cross-site data transfer risks and achieving compliance with the General Data Protection Regulation and Act on the Protection of Personal Information data minimization principles. Complementary privacy-preserving measures—including differential privacy and secure aggregation protocols—are recommended to guard against gradient inversion attacks that can partially reconstruct training data from shared gradients. UltraFedFM demonstrated this approach at scale, pre-training across 16 institutions in nine countries using over one million unlabeled images to achieve an AUROC of 0.927 for disease classification and a Dice coefficient of 0.878 for lesion segmentation (*9*). For a multi-center trial using a PCCP-governed AI endpoint, federated learning provides the infrastructure through which algorithm updates agreed upon in the PCCP can be implemented across all sites without requiring patient data transfer—converting what would otherwise be a regulatory and logistical obstacle into a routine, pre-authorized model maintenance cycle.

6. Future directions and challenges

6.1. Validation standards

The most pressing unresolved challenge for AI-assisted ultrasound in clinical trials is the absence of agreed minimum standards for external validity. The evidence hierarchy described in Section 3.2 provides a conceptual framework, but trial sponsors currently have no single authoritative checklist specifying what degree of external

validation is sufficient before an AI measurement algorithm is adopted to ascertain a primary trial endpoint. Translating STARD-AI and PROBAST+AI principles into a concise, endpoint-specific validation template could lower the barrier to regulatory acceptance and reduce redundant validation work across independent trial programs (*14,15*).

6.2. Data infrastructure and foundation models

The federated foundation model paradigm exemplified by UltraFedFM (*9*) suggests a potentially important direction for ultrasound AI infrastructure. Rather than developing a separate model de novo for each trial, in the future AI-assisted ultrasound may increasingly be adapted from large pre-trained models that are subsequently fine-tuned for specific clinical or trial applications and which are then used to ascertain endpoints. If validated across endpoint-specific use cases, such an approach could reduce the amount of task-specific labelled data required and may improve cross-site robustness by capturing a broader range of scanner, operator, and population variability during pre-training. However, whether these advantages translate consistently to regulated trial endpoints still needs to be established. Realizing this potential would likely require expansion of federated training networks, with broader representation of community hospitals and LMICs. More diverse participation could help reduce the risk that foundation models reflect predominantly tertiary-center patient populations, workflows, or equipment characteristics.

The public data ecosystem also requires investment. The SonoDQS catalog identified 72 public datasets as of late 2025, but fetal and prostate ultrasound remain substantially underrepresented relative to the frequency of their clinical use (*13*). Enhancing coordinated data-sharing initiatives—and particularly those that include standardized acquisition metadata, demographic descriptors, and reference-standard labels—could facilitate both more reproducible algorithm development and more credible independent benchmarking.

6.3. Limitations of this review

Several limitations should be acknowledged. First, this is a narrative review and does not claim to represent a systematic or exhaustive survey of the literature; papers were selected to illustrate key methodological and regulatory principles, and the evidence presented may not reflect the full breadth of published findings. Second, the clinical validation evidence is weighted toward AI-assisted echocardiography, reflecting the current state of the field: rigorous multi-center prospective studies and RCTs mostly involve cardiac imaging, while evidence levels are lower for fetal, hepatic, musculoskeletal, and vascular ultrasound. Conclusions about the regulatory and methodological readiness of AI endpoints in non-cardiac domains should therefore be interpreted as anticipatory rather than established. Third, this review draws primarily on English-language literature and published regulatory guidance from the United States, European Union, and Japan; evidence and frameworks from other jurisdictions—including China, South Korea, and Brazil, each of which has issued AI medical device guidance—were not systematically covered. Fourth, the regulatory landscape is changing rapidly: specific timelines, guidance documents, and device inventories cited here reflect the state of affairs as of early 2026, and readers should verify their current status through primary regulatory sources before relying on these details for submission purposes.

7. Conclusion

AI-assisted ultrasound is advancing from research proof-of-concept toward operational use as a clinical trial tool, with a growing body of prospective evidence and formalized guidance supporting this transformation. The three evidence domains reviewed here—automated endpoint measurement, AI-guided ultrasound acquisition in decentralized settings, and regulatory and methodological readiness—each now rest on more than speculative promise, but routine adoption is limited in most non-cardiac ultrasound applications.

For investigators designing trials with ultrasound-based endpoints, the practical implications are increasingly clear. Automated measurement algorithms should be validated against a pre-specified equivalence or performance criterion before using them to ascertain primary endpoints, and subgroup performance and domain shift analyses should be reported transparently in accordance with frameworks such as STARD-AI or PROBAST+AI. AI-guided ultrasound acquisition protocols should be evaluated in the intended operator population - not merely in trained sonographers - before adoption in DCT designs. Regulatory interactions should consider change-management planning from the outset so that future algorithm modification is handled through a prospectively defined process rather than an avoidable protocol-amendment crisis.

For regulators and guideline developers, the most productive near-term investment is in harmonization: of cross-vendor benchmark standards, of endpoint validation templates, and of data sharing norms that shape public ultrasound resources to meet clinical need. These convergences will not occur spontaneously; they require coordinated action across the research, industry, and regulatory communities that this review addresses collectively.

Funding: This work was supported by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (24K14216).

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

References

1. Shen YT, Chen L, Yue WW, Xu HX. Artificial intelligence in ultrasound. Eur J Radiol. 2021; 139:109717.
2. Thavendiranathan P, Grant AD, Negishi T, Plana JC, Popović ZB, Marwick TH. Reproducibility of echocardiographic techniques for sequential assessment of left ventricular ejection fraction and volumes: Application to patients undergoing cancer chemotherapy. J Am Coll Cardiol. 2013; 61:77-84.
3. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, Heidenreich PA, Harrington RA, Liang DH, Ashley EA, Zou JY. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020; 580:252-256.
4. He B, Kwan AC, Cho JH, *et al*. Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature. 2023; 616:520-524.
5. van Sloun RJG, Cohen R, Eldar YC. Deep learning in ultrasound imaging. Proc IEEE. 2020; 108:11-29.
6. Trost B, Rodrigues L, Ong C, *et al*. Artificial intelligence empowers novice users to acquire diagnostic-quality echocardiography. JACC Adv. 2025; 4:102005.
7. U.S. Food and Drug Administration. Conducting Clinical Trials With Decentralized Elements: Guidance for Industry, Investigators, and Other Interested Parties. FDA; September 2024. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/conducting-clinical-trials-decentralized-elements* (accessed April 1, 2026).
8. Vafaeezadeh M, Behnam H, Gifani P. Ultrasound image analysis with vision transformers — Review. Diagnostics. 2024; 14:542.
9. Jiang Y, Feng CM, Ren J, *et al.* From pretraining to privacy: Federated ultrasound foundation model with self-supervised learning. NPJ Digit Med. 2025; 8:714.
10. Tromp J, Bauer D, Claggett BL, *et al*. A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. Nat Commun. 2022; 13:6776.
11. UUSIC25 Challenge Consortium. Diagnostic performance of universal-learning ultrasound AI across multiple organs and tasks. arXiv. 2025. arXiv:2512.17279.
12. Mendez M, Jafarpisheh N, Demello S, Lee C, Dang M, Tyrrell PN. Evaluating the impact of sex bias on AI models in musculoskeletal ultrasound of joint recess distension. PLoS One. 2025; 20:e0332716.
13. Alsharid M, Guo X, Men Q, *et al*. On the public dissemination and open sourcing of ultrasound resources, datasets and deep learning models. NPJ Digit Med. 2025; 8:777.
14. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, Golub RM, Denniston AK, Shetty S, Moher D, Bossuyt PM, Darzi A, Ashrafian H. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025; 31:3283-3289.
15. Moons KGM, Damen JAA, Kaul T, *et al*. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025; 388:e082505.
16. Vasey B, Nagendran M, Campbell B, *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022; 28:924-933.
17. Lekadir K, Frangi AF, Porras AR, *et al.* FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025; 388:e081554.
18. Baloescu C, McNaughton CD, Arntfield R, *et al*. Artificial intelligence-guided lung ultrasound by nonexperts. JAMA Cardiol. 2025; 10:245-253.
19. Speranza G, Mischkewitz S, Al-Noor F, Kainz B. Value of clinical review for AI-guided deep vein thrombosis diagnosis with ultrasound imaging by non-expert operators. NPJ Digit Med. 2025; 8:135.
20. U.S. Food and Drug Administration. Clinical Trial Imaging Endpoint Process Standards: Guidance for Industry. FDA; 2018. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-imaging-endpoint-process-standards-guidance-industry* (accessed April 1, 2026).
21. Wong A, Roslan NL, McDonald R, *et al*. Clinical obstacles to machine-learning POCUS adoption and system-wide AI implementation (The COMPASS-AI survey). Ultrasound J. 2025;17:32.
22. Kim S, Fischetti C, Guy M, Hsu E, Fox J, Young SD. Artificial intelligence (AI) applications for point of care ultrasound (POCUS) in low-resource settings: A scoping review. Diagnostics. 2024; 14:1669.
23. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and FDA Staff. FDA; December 2024. *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence* (accessed April 1, 2026).
24. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024; L 2024/1689. *https://eur-lex.europa.eu/eli/reg/2024/1689/oj* (accessed April 1, 2026).
25. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. FDA; October 2021. *https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles* (accessed April 1, 2026).
26. Stringer JSA, Pokaprakarn T, Prieto JC, *et al*. Diagnostic accuracy of an integrated AI tool to estimate gestational age from blind ultrasound sweeps. JAMA. 2024; 332:649-657.
27. Slimani S, Hounka S, Mahmoudi A, *et al.* Fetal biometry and amniotic fluid volume assessment end-to-end automation using deep learning. Nat Commun. 2023; 14:7047.
28. Zhang L, Zhao X, Song L, *et al*. Paired liver-spleen high-frequency ultrasound deep learning network for full-stage liver fibrosis classification and clinical benefit compared with 2D-SWE in chronic hepatitis B cohort: A prospective

multicenter study. J Gastroenterol. 2026; 61:184-194.
29. Ammitzbøll-Danielsen M, Østergaard M, Tamm L, Terslev L. Diagnostic performance and reliability of robotic ultrasonography and artificial intelligence-driven synovitis assessment in rheumatoid arthritis: Results from the Controlled ARTHUR Trial. RMD Open. 2025; 11:e006099.
30. International Medical Device Regulators Forum. Good machine learning practice for medical device development: Guiding principles. IMDRF; 2025. *https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles* (accessed April 1, 2026).
31. Gomes RG, Vwalika B, Lee C, *et al.* A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. Commun Med. 2022; 2:128.
32. Chiu IM, Chen TY, Zheng YC, Lin XH, Cheng FJ, Ouyang D, Cheng CY. Prospective clinical evaluation of deep learning for ultrasonographic screening of abdominal aortic aneurysms. NPJ Digit Med. 2024; 7:282.
33. U.S. Food and Drug Administration. Artificial Intelligence-enabled Medical Devices. *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices* (accessed April 1, 2026).
34. Pharmaceuticals and Medical Devices Agency. Regulations and Approval/Certification of Medical Devices in Japan. *https://www.pmda.go.jp/english/review-services/reviews/0004.html* (accessed April 1, 2026).
35. Pharmaceuticals and Medical Devices Agency. Software as a Medical Device (SaMD). *https://www.pmda.go.jp/english/review-services/reviews/0009.html* (accessed April 1, 2026).
36. Pharmaceuticals and Medical Devices Agency. SAKIGAKE Designation System. *https://www.pmda.go.jp/english/review-services/reviews/advanced-efforts/0001.html* (accessed April 1, 2026).

Received March 10, 2026; Revised April 12, 2026; Accepted April 18, 2026.

**Address correspondence to:*
Kenji Karako, Department of Surgery, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, Japan 113-8655.
E-mail: tri.leafs@gmail.com

Released online in J-STAGE as advance publication April 20, 2026.