npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-03155-7

A five-phase evaluation framework for diagnostic and predictive medical artificial intelligence

Check for updates

Zichen Ye[1] ✉, Yue Chen[2], Xuefeng Huang[3], Manman Chen[3], Wanzhou Wang[4,5,6], He Zhu[1], Keyu Han[1], Jiahui Wang[1], Qu Lu[1], Yuankai Zhao[3], Yimin Qu[3] ✉, Guanghan Gao[7] ✉ & Yu Jiang[1] ✉

Artificial intelligence (AI) has advanced rapidly across diagnostic, prognostic, and clinical decision-support applications, yet the pathway from laboratory performance to demonstrable clinical benefit remains fragmented and inconsistently defined. Existing evaluations rely heavily on retrospective testing and algorithm-centric metrics, while current guidelines emphasize reporting standards rather than specifying validation across stages of model maturity. This study proposes a five-phase evaluation framework for medical AI, supported by a dynamic evaluation architecture reflecting the nonlinear, iterative nature of AI systems. The framework integrates technical validation, operational robustness validation, controlled interaction validation, clinical evidence validation, and real-world integration validation, while incorporating phase-gating criteria and local and systemic fall-back triggers. These mechanisms enable re-entry into earlier phases based on drift, version updates, or safety signals, and accommodate parallel activities such as implementation research informing clinical trials. By systematically mapping multicenter external validation, shadow-mode testing, human–AI comparison and cooperation studies, randomized controlled trials, real-world evaluations, and adaptive designs into a coherent lifecycle pathway, the framework addresses persistent gaps between laboratory performance and clinical benefit. It provides researchers, clinical institutions, and regulators with an operational, scalable approach aligned with evolving regulatory expectations, supporting trustworthy, ethically aligned, and lifecycle-based evidence generation for medical AI systems.

Artificial intelligence (AI) has advanced rapidly across diagnostic, prognostic, and clinical decision-support applications, [1–3] including imaging[4–6] and pathology[7–9], yet the pathway from laboratory performance to demonstrable clinical benefit remains methodologically underdeveloped[10–12]. Most current studies rely on retrospective testing[13], focusing on algorithmic metrics under controlled conditions. However, clinical practice is primarily concerned with whether AI improves diagnostic workflows and patient outcomes[14,15]. Meanwhile, existing reporting guidelines such as DECIDE-AI[16], TRIPOD + AI[17], and STARD-AI[18] mainly address documentation standards, whereas broader translation frameworks and technology-readiness scales operate at macro-level engineering or regulatory milestones[19–21]. Neither provides prescriptive guidance on which clinical study designs to use at different maturity stages or how validation should progress in a structured manner. This methodological gap has led to a persistent "evaluation mismatch"[22,23]: on one hand, many AI systems proceed to costly clinical trials before achieving sufficient stability; on the other hand, some systems remain confined to offline comparisons even when clinical benefit needs to be demonstrated.

To address this gap, we propose the concept of a medical AI clinical evaluation gradient framework, which views validation as a continuous,

[1]School of Health Policy and Management, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. [2]Division of cancer epidemiology, German cancer research center. Im Neuenheimer Feld 581, Heidelberg, Germany. [3]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. [4]Institute of Medical Technology, Peking University Health Science Center, Beijing, China. [5]National Institute of Health Data Science at Peking University, Beijing, China. [6]Center for Digital Health and Artificial Intelligence, Peking University First Hospital, Beijing, China. [7]Department of Orthopedics, Beijing Luhe Hospital, Capital Medical University, Beijing, China. ✉e-mail: ye18700579760@163.com; quyimin@sph.pumc.edu.cn; lxxx-gaoguanghan@163.com; jiangyu@pumc.edu.cn

nonlinear lifecycle rather than a one-off performance assessment. The framework encompasses data robustness, engineering stability, human–AI interaction, clinical evidence, and real-world integration. The model development stage serves as the foundation, establishing initial performance boundaries before formal evaluation begins. Crucially, although the framework depicts model development as its foundational layer, it presupposes an upstream, problem-led clinical-need assessment to ensure that an AI intervention is epidemiologically and operationally justified.

Building on this principle, we develop a lifecycle-based evaluation framework consisting of five-phase evaluation framework supported by an overarching dynamic evaluation architecture that explicitly accommodates nonlinear parallelism (Fig. 1). The framework systematically integrates eight core methodological approaches and eighteen study designs (Figs. 1 and 2), enabling researchers to identify the maturity stage of their system and select appropriate study designs accordingly. In doing so, it shifts AI research from algorithm-centric performance comparisons toward genuine clinical evidence generation, ensuring that AI technologies can be deployed safely, effectively, and with meaningful value in real healthcare settings.

Although conceptually general, the framework is primarily applicable to diagnostic and prognostic medical AI systems, particularly those entering clinical use through the SaMD pathway. These systems typically adopt a frozen-model paradigm during evaluation, and their evidence-generation trajectory aligns with that of traditional diagnostic devices. At the same time, the five-phase framework is intended as a flexible evaluation gradient rather than a rigid sequence; the appropriate pathway depends on intended use, risk class, regulatory designation, and clinical context. The dynamic evaluation architecture further outlines how continually learning models and non-diagnostic AI applications may selectively implement or adapt components of the framework. The framework also aligns with emerging global governance expectations, including robustness, fairness, transparency, and lifecycle risk management emphasized in the OECD AI Principles[24] and the EU AI Act[25].

Importantly, the framework accommodates medical AI systems with different levels of algorithmic autonomy. Assistive or human-in-the-loop systems require explicit evaluation of human–AI cognitive coupling, which is the central focus of Phase 3. Fully autonomous systems may have limited need for the human–AI interaction components of Phase 3 but require higher evidentiary standards in Phases 1, 2, and 4, as they operate without a clinician safety net. With this scope and autonomy matrix established, the following sections outline the core methodologies across the five validation phases, beginning with the foundational requirement of technical validation.

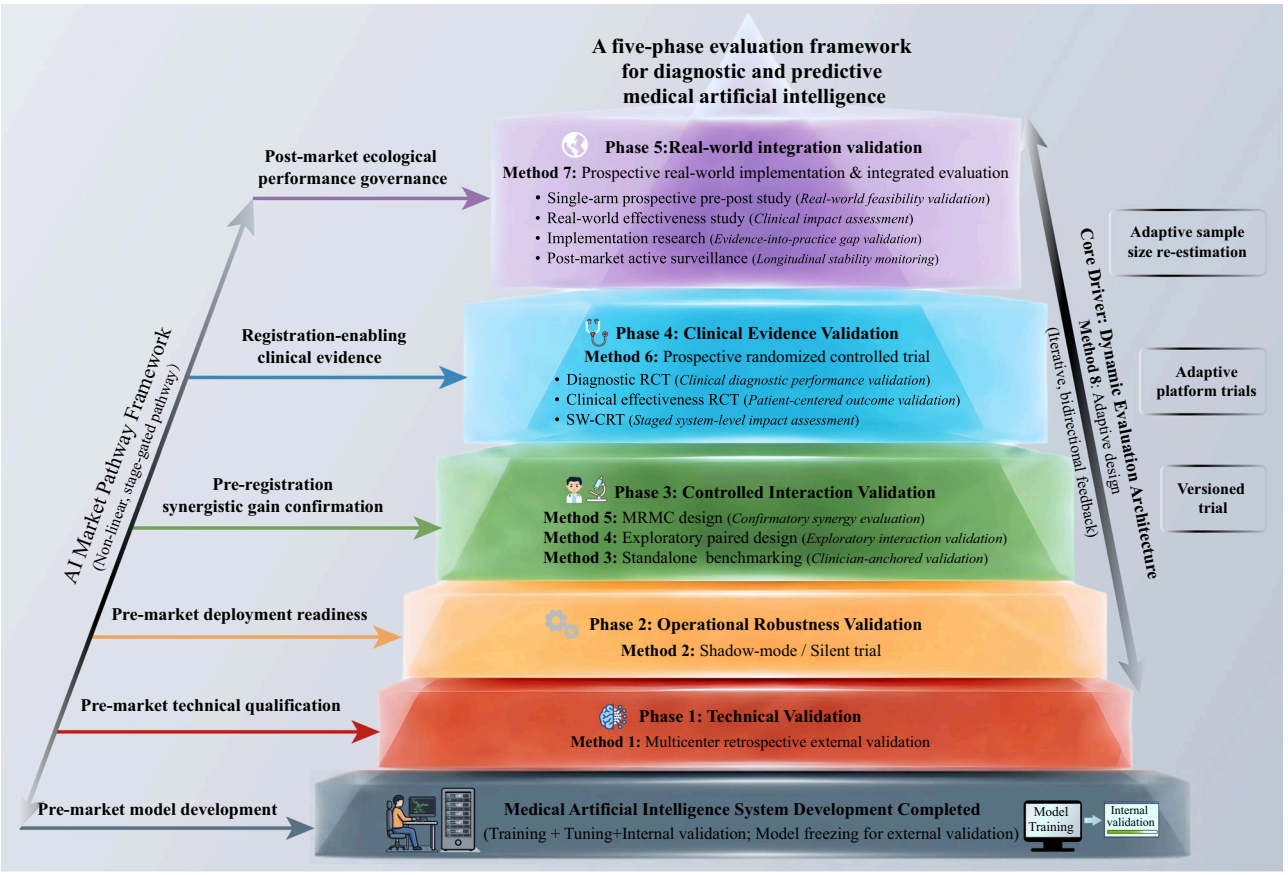


Fig. 1 | A five-phase evaluation framework for diagnostic and predictive medical artificial intelligence. The framework outlines a full-lifecycle evaluation pathway, extending from model development to long-term real-world integration. The model development stage, encompassing training, tuning, and internal validation, forms the technical foundation from which formal evaluation begins. Importantly, the framework presupposes an upstream, problem-led clinical-need assessment prior to any algorithmic development, ensuring that evaluation proceeds only after a clinically meaningful use-case has been established. The model evaluation stage comprises five validation phases: (1) Phase 1: technical validation; (2) Phase 2: operational robustness validation; (3) Phase 3: controlled interaction validation; (4) Phase 4: clinical evidence validation; and (5) Phase 5: real-world integration validation. The right side of the figure presents the dynamic evaluation architecture, which serves as the core driver of the entire framework. The left side depicts the AI market pathway framework, illustrating the progression from pre-market model development and technical qualification to registration-enabling clinical trials and post-market system-level performance governance. Note: The five phases, market pathway, and dynamic evaluation architecture do not represent a linear sequence. Bidirectional transitions, parallel evaluation activities, and phase re-entry may occur throughout the medical AI lifecycle. AI artificial intelligence, RS reference standard, MRMC multi-reader multi-case, RCT randomized controlled trial, SW-CRT stepped-wedge cluster randomized controlled trial.

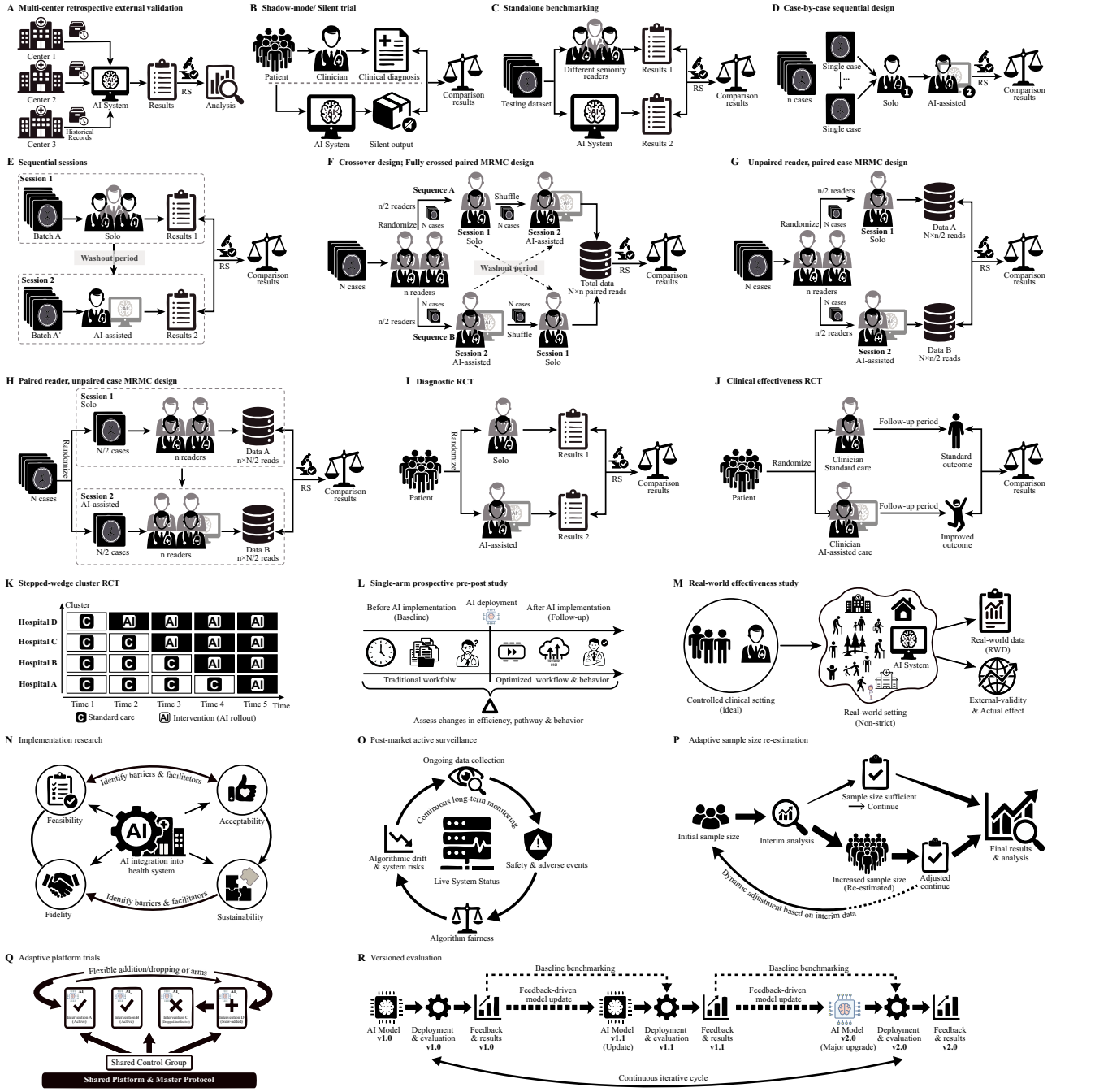


Fig. 2 | Eighteen study designs mapped to the five-phase evaluation framework and its dynamic evaluation architecture for diagnostic and predictive medical artificial intelligence. Method 1: **A** Multi-center retrospective external validation; Method 2: **B** Shadow-mode/silent trial; Method 3: **C** Standalone benchmarking; Method 4. Exploratory paired design: **D** Case-by-case sequential design; **E** Sequential sessions; **F** Crossover design; Method 5. Multi-reader multi-case: **F** Fully crossed paired MRMC design; **G** Unpaired reader, paired case MRMC design; **H** Paired reader, unpaired case MRMC design; Method 6. Prospective randomized controlled trial: **I** Diagnostic RCT; **J** Clinical effectiveness RCT; **K** Stepped-wedge cluster RCT; Method 7. Prospective real-world implementation and integrated evaluation: **L** Single-arm prospective pre-post study; **M** Real-world effectiveness study; **N** Implementation research; **O** Post-market active surveillance; Method 8. Adaptive design: **P** Adaptive sample size re-estimation; **Q** Adaptive platform trials; **R** Versioned evaluation. AI artificial intelligence, MRMC multi-reader multi-case, RCT randomized controlled trial, RS reference standard.

Results

Phase 1. Technical validation: establishing data robustness

Rationale and positioning. Technical validation aims to assess whether a frozen medical AI model can maintain stable performance when applied to fully independent and clinically heterogeneous datasets. Rather than further optimizing model architecture or parameters, this stage assesses generalizability across institutions, data-generation or acquisition processes, and patient populations. The goal is to establish a data robustness baseline under idealized data flows, thereby creating the prerequisite conditions for subsequent operational robustness and clinical evaluation. Multi-center retrospective external validation (Method 1) represents the first methodological gateway for medical AI to enter the clinical evaluation pathway[26]. Its core question focuses on whether a model can remain reliable once decoupled from its development

environment and tested on independent, clinically representative data. To ensure methodological transparency, external validation studies in this phase must align with task-specific international standards: studies evaluating the standalone diagnostic accuracy of an AI tool should adhere to the STARD-AI statement[18], whereas those validating multivariable individual risk prediction models should strictly follow the TRIPOD + AI guidelines[17].

Key design elements. Rigorous technical validation requires strict governance of both the study population and the data pipeline. Crucially, prior to any data access or analysis, a formal study protocol and a detailed statistical analysis plan (SAP) must be explicitly prespecified. This phase pre-specification ensures that primary endpoints, subgroup stratifications, and success criteria are not retroactively fitted to the data. Before evaluation, the model version, parameters, and decision thresholds must be locked, and no post-hoc tuning is permitted[27]; any modification constitutes a new model and must be re-evaluated on an entirely new validation set. Dataset selection should follow a diagnostic cohort design, using consecutive enrollment or random sampling to preserve low-quality images and complex comorbidities rather than selecting "typical cases", which is essential for detecting and avoiding spectrum bias but remains challenging in practice[28,29]. The validation set must be fully independent from the development data in temporal (different acquisition periods), geographic (different institutions), or both[30,31]. Inclusion of multiple centers (typically ≥3) is recommended to capture natural heterogeneity in data sources, acquisition processes, and populations, thereby probing the model's generalizability limits[27,32,33]. To ensure that technical validation reflects genuine robustness rather than overfitting to a single-site distribution, Phase 1 must explicitly characterize and address structural sources of heterogeneity across participating centers. This includes documenting and mitigating case-mix differences, acquisition-process variation, prevalence shifts, and any site-level differences in data-generation workflows.

Reference standard construction should maintain strict separation between the algorithm team and the annotation team. For retrospective studies, pathology confirmation or expert consensus should be used[34], and delayed follow-up (e.g., 1–3 years) may be necessary to correct verification bias[35,36] and decouple the ground truth from initial clinical diagnosis[35,37]. To combat structural validity problems caused by site heterogeneity, rigorous reference standard harmonization is mandatory to ensure uniform ground-truth definition across diverse sites. Furthermore, if a single global threshold yields suboptimal calibration across varying clinical environments, guidance for local threshold recalibration should be established and documented. If validation results fall short of expectations, in-place parameter adjustments are prohibited. Instead, systematic failure analysis should be conducted to identify vulnerable scenarios, such as specific devices, pathological severity, or data-quality anomalies, and these findings should be fed back into model development. Any updated version must be treated as a new model and re-evaluated on a new, independent third-party dataset to preserve the integrity and independence of the evaluation process. Finally, the evaluation must include formal AI-by-site interaction testing alongside stratified robustness checks across key heterogeneity factors. These measures ensure that technical validation probes the model's true stability across heterogeneous inputs rather than relying on a single dominant data source.

Sample size determination should be based on the target sensitivity or specificity, desired statistical power, and significance level. Additional inflation is often required to account for center-level clustering and potential data drift, ensuring that statistical inference remains robust across heterogeneous multi-center settings (Supplementary Material 1)[38,39].

Evaluation components. In this phase, overall discrimination metrics such as the area under the receiver operating characteristic curve (AUROC) or the area under the precision-recall curve should be reported alongside sensitivity at a prespecified specificity, as well as the corresponding positive and negative predictive values (PPV/NPV)[40]. Beyond discrimination, probabilistic models must report calibration metrics (e.g., Brier score, expected calibration error [ECE], and reliability plots) together with validated uncertainty estimates to evaluate prediction reliability, particularly for heterogeneous data, edge cases, and out-of-distribution inputs. Additionally, decision-curve analysis (DCA) should be incorporated to estimate the potential net clinical benefit across relevant risk thresholds[41]. Key subgroup analyzes, stratified by pathological severity, data-source heterogeneity, acquisition processes, and demographics, alongside formal AI-by-site interaction testing, are essential to avoid masking failure modes that may be hidden by "average performance"[40,42]. Subgroup specification must be driven by clinical relevance rather than modality-specific conventions. The imaging-related subgroup examples provided here are illustrative rather than prescriptive; actual subgroup definitions should be determined by clinical experts based on disease context, patient heterogeneity, and the intended clinical use.

Strengths and limitations. The core value of method lies in using heterogeneous clinical data from independent multiple sites to test the upper bounds of model performance and the limits of robustness across settings[43–45], and to identify overfitting, spectrum bias, and inter-institutional variability that single-center studies cannot reveal[43,44]. The approach provides the safety boundary required for follow-up research. However, because it relies on historical and often "cleaned" data, 28, 29 it cannot assess human–AI interaction effects, may not reflect contemporary clinical practice, and is susceptible to variability in data quality, annotation practices, and selection bias across centers[40,42]. Consequently, it should serve as the starting point rather than the endpoint of the evidence chain, establishing the model's data robustness baseline under idealized data flows, rather than demonstrating clinical effectiveness.

Phase 2. Operational robustness validation: engineering stability and stream stress testing

Rationale and positioning. The operational robustness validation phase evaluates the engineering stability and data-stream adaptability of a medical AI system within real hospital workflows. Unlike technical validation, which focuses on data robustness in static datasets, this phase examines how the model performs when exposed to raw clinical data streams. The core question is whether the system can process reliably and stably on real-world data within a clinical environment, before influencing clinical decision-making.

Shadow-mode or silent trial (Method 2) represents the core and most practical study design for this phase. The model runs in real time in the background without displaying outputs to clinicians, allowing comprehensive assessment of its behavior under authentic data flow conditions, including data preprocessing, inference latency, failure rates, and engineering bottlenecks[46–48]. The purpose of this phase is to establish an engineering safety boundary before human–AI interaction studies or clinical trials, ensuring that the system can "operate, operate stably, and operate end-to-end" in real clinical environments.

Key design elements. Crucially, prior to initiating shadow-mode deployment, a formal monitoring protocol or a versioned analysis charter must be explicitly prespecified. The model must be deployed within clinical data pathways to ensure full alignment of data formats, acquisition processes, and metadata structures with routine clinical workflows[46]. Shadow-mode deployment should encompass all clinical cases, rather than only "readable" or "typical" cases, to expose potential failure modes arising from low-quality images, atypical protocols, or rare findings.

At the engineering level, key indicators of the data pipeline, including data ingestion, preprocessing, inference latency, resource utilization, and batch-processing stability, should be systematically monitored, with all abnormal events logged together with their triggers and frequencies[46–48]. In settings involving multiple data sources, acquisition processes, or site-level

workflows, the model's stability and agreement under varying input distributions must be evaluated to identify risks of data drift and model drift[26,46]. A comprehensive logging and error-tracking system is required to ensure that every failure can be located, reproduced, and explained[46].

Organizationally, the algorithm team and clinical team should remain separated, and clinicians must not be exposed to model outputs to avoid automation bias[49]. Issues identified during this phase should be addressed through engineering fixes or data governance rather than parameter tuning or retraining; if model updates are necessary, they should be treated as a new version and re-evaluated through a new shadow-mode cycle.

Sample size estimation for shadow-mode testing prioritizes the confirmation of engineering robustness over conventional hypothesis testing. Consequently, the scale is determined by a synthesis of agreement, stability, and safety (Supplementary Material 2). First, Cohen's kappa or intraclass correlation coefficients (ICC)[50,51] can be used to estimate the number of cases required to ensure that AI-clinician agreement meets predefined safety thresholds. Second, confidence interval width method or single-proportion formulas[52–54] are used to estimate system failure rates to ensure engineering robustness attains the targeted precision. Third, the "rule of three"[55–57] or binomial distribution[58,59] can be applied to estimate the upper bound of rare critical errors or to determine the sample size needed to capture at least one such event. Given the dynamic nature of real-world data streams and potential data loss[46,60], a 10–20% inflation over the theoretical estimate is typically recommended to ensure adequate retention, rather than to compensate for underlying structural site heterogeneity.

Evaluation components. The focus of this phase has expanded from single-dimension diagnostic performance to a multidimensional evaluation framework encompassing model performance, system robustness, diagnostic agreement, and data-stream adaptability[46,47,61–63]. In addition to Phase 1 metrics (e.g., AUC, sensitivity, specificity, and calibration), Phase 2 emphasizes real-time operational capability within hospital environments. This includes engineering stability indicators such as failure rate, inference latency, system throughput, and service availability, alongside resource-load characteristics such as compute cost, energy consumption, network bandwidth, and performance under concurrent load (e.g., latency at scale, batch throughput, and compute-resource ceilings). These factors jointly determine whether the system can be deployed sustainably across departments and devices in real-time, high-volume clinical workflows while accounting for long-term infrastructure and environmental impact.

Diagnostic agreement between AI outputs and the reference standard should be quantified using measures such as Cohen's kappa or ICC. In addition, Phase 2 requires explicit monitoring of calibration and predictive uncertainty under real-world data streams. Calibration drift should be assessed using reliability plots, Brier scores, and ECE[41,64], while uncertainty estimates should be evaluated for their ability to characterize prediction reliability and identify potential out-of-distribution inputs without influencing clinical decisions. Furthermore, data drift, device heterogeneity, and potential biases should be evaluated through drift-detection metrics, bias audits, and fairness assessments across predefined clinical subgroups. However, these assessments represent baseline characterization rather than longitudinal monitoring; the objective is to establish the system's initial robustness profile when first exposed to real-world data streams without influencing clinical decisions. Continuous, time-varying drift and fairness monitoring is reserved for Phase 5, where post-market surveillance governs long-term performance stability.

Strengths and limitations. The primary value of the phase lies in prospectively and non-interventionally validating the engineering stability and data-stream adaptability of a medical AI system under clinical workflows[46–48]. By receiving raw clinical inputs in real time, the model avoids the selection bias inherent to retrospectively curated datasets and more accurately reflects performance under authentic input distributions. This phase also enables high-fidelity engineering validation by fully reproducing hospital data pathways, including clinical data-transmission pipelines, network latency, protocol variability, and metadata heterogeneity, thereby revealing engineering vulnerabilities. Because model outputs do not influence clinical decisions, the ethical risk is low[26,49], and the frozen outputs can be compared with clinician diagnoses or reference standards at study completion to generate high-quality performance evidence[49].

However, this phase does not involve clinician behavior[48] and therefore cannot assess human–AI interaction effects, automation bias, or workflow changes, nor can it directly reflect impacts on diagnostic efficiency or patient outcomes. Furthermore, implementation is resource-intensive, demanding robust hospital information technology infrastructure, data governance, and privacy protection. As an engineering assessment, this phase should be regarded as an engineering threshold within the evidence chain, establishing system stability under real data flows rather than demonstrating clinical effectiveness[46–48].

Phase 3. Controlled interaction validation: human–AI interaction and cognitive assessment

The controlled interaction phase represents the critical transition in medical AI from non-interventional validation to a clinical decision-support application. Unlike earlier stages that emphasize standalone model performance, this phase focuses on systematically evaluating human–AI interaction effects in controlled, simulated clinical environments, with particular attention to changes in clinicians' diagnostic direction, confidence, and decision efficiency following exposure to AI outputs[26,48].

To avoid the ethical and safety risks of testing human–AI interaction directly in real clinical settings, clinicians are brought into a controlled experimental environment where fatigue, external distractions, and workflow pressures are minimized. This allows precise measurement of the AI's net cognitive benefit and clarification of its role in the decision-making chain, whether it functions as a "parallel peer" or as a "cognitive enhancer" providing additional information. Depending on the study objective and depth of assessment, Phase 3 comprises three complementary methodological pathways that together form the core framework for evaluating cognitive contribution and human–AI collaboration. Crucially, across all three pathways, a formal study protocol and SAP must be explicitly prespecified prior to any experimental testing. To standardize the evidence-generation process for clinical decision support and human factors, studies in this phase should be guided primarily by the DECIDE-AI guidelines[16]. Additionally, if the primary objective includes quantifying the final collaborative diagnostic accuracy of the unified human–AI team, the reporting framework should be complementarily structured around the STARD-AI standard[18].

Method 3. Standalone benchmarking

Rationale and positioning. Standalone benchmarking aims to determine the "performance position" of a medical AI system by conducting blinded, parallel comparisons between the model and clinicians of different seniority levels (junior, mid-career, or senior) on the same standardized case set[65,66]. This method addresses a core scientific question: To which level of clinical expertise is the AI's independent diagnostic ability equivalent? This positioning is essential for evaluating the feasibility of using AI as a second reader, a task-shifting tool in resource-limited settings, or an expert-support tool[65,67]. By enabling a pure comparison of cognitive diagnostic performance, it provides the baseline for subsequent human–AI collaboration studies and serves as the foundational prerequisite for interaction evaluation.

Key design elements. The design of standalone benchmarking centers on constructing a high-quality, reproducible, and clinically representative standardized dataset, and ensuring a blinded, parallel evaluation between AI and clinicians. The datasets should capture natural heterogeneity in disease spectrum, pathological severity, input-quality variation, data-source heterogeneity, and patient characteristics[68,69], with reference standards established by an independent expert panel or

pathology rather than initial clinical reads[28,29]. Clinicians of varying seniority should be recruited to create a graded reference framework that pinpoints the AI within the clinical decision pathway[28,29]. The evaluation adheres to the principles of blinding, randomization, and parallel assessment: blinding prevents information contamination; randomization mitigates order and learning effects; and parallel evaluation ensures that AI and clinicians diagnose identical information independently, yielding highly comparable data.

Sample size is primarily driven by the number of cases and is typically determined through independent (non-paired) comparisons of AUC[70,71] or differences in sensitivity and specificity[72,73] between AI and clinicians. Although there is no strict minimum number of participating readers, it is generally recommended to include at least three clinicians per seniority tier to mitigate individual variability and ensure robustness of the findings[29,74] (Supplementary Material 3).

Evaluation components. The evaluation framework must capture both diagnostic performance and inter-reader variability. Core metrics include AUROC, sensitivity, specificity, and PPV/NPV at clinically relevant thresholds[28,29]. For classification tasks, agreement between AI and expert readers can be quantified using Cohen's kappa or weighted kappa[63,75]; for continuous scoring tasks, ICC or Spearman correlation may be used. For probabilistic prediction or triage-oriented systems, additional assessment of calibration and uncertainty estimation performance may be informative, using approaches such as reliability plots, Brier score, ECE, and appropriate empirical validation of uncertainty estimates.

Strengths and limitations. The strength of standalone benchmarking lies in its ability to generate a pure comparison of cognitive diagnostic performance between AI and clinicians of different seniority, thereby defining the AI's performance position and providing direct evidence for its potential role within the clinical pathway[29,65]. This approach can reveal the model's strengths and limitations across specific lesion types, image qualities, and case difficulties, and establishes the baseline for subsequent human–AI collaboration studies. However, its limitations stem from the inability to assess automation bias, workflow effects, or the influence of AI on clinicians' behavior. In addition, the generalizability of findings depends heavily on the quality of the datasets, and inappropriate case selection may introduce spectrum bias or lead to over- or under-estimation of the model's true capability[69]. Therefore, it should be regarded as a cognitive-level positioning tool that characterizes the AI's independent diagnostic ability, rather than a basis for inferring its ultimate clinical impact. Clinician-seniority benchmarking is most appropriate for interpretive specialties where diagnostic performance exhibits clear seniority gradients. For generalist or decision-support tasks, and for generative or agentic AI systems, alternative benchmarking strategies are required, as seniority-based comparisons may not meaningfully reflect task-specific competence. It marks the starting point of Phase 3, not its endpoint.

Method 4. Exploratory paired design

Rationale and positioning. The exploratory paired design aims to evaluate the direct impact of AI on clinicians' diagnostic behavior in a controlled environment, focusing on how the same clinician's diagnostic direction, confidence, or efficiency changes with versus without AI assistance. Using a self-controlled framework, each clinician interprets the same case set under both "AI-assisted" and "solo" conditions, enabling the detection of directional effects on cognitive decision-making with minimal sample size[76,77]. To accommodate different study objectives and memory-effect constraints, several structured variants of the paired design have been developed, including case-by-case sequential, sequential sessions, and crossover designs[76], each with distinct features summarized in Box 1.

Box 1. | Overview of paired designs and their variants

1. Case-by-case sequential design. It requires each reader to complete two consecutive steps for every case: an initial solo diagnosis with the result locked, followed immediately by an AI-assisted diagnosis of the same case before moving to the next one (Fig. 3A). Key features: (1) Operational simplicity. The workflow is compact and requires no separate reading sessions, making it the least resource-intensive of all paired designs. (2) Risk of learning effects. Because readers may gradually internalize AI patterns after repeatedly viewing AI outputs, artificially improving subsequent solo performance and thereby underestimating AI's true incremental benefit. (3) Limited applicability. This design is suitable only for evaluating AI in a second-reader workflow (human-first, AI-second)[67,83] and cannot assess AI-only or parallel-assistance modes.

2. Sequential-session design. In this design, the reading process is divided into two fully independent sessions: all cases are first diagnosed without AI, followed, after an adequate washout period, by a second session in which the same cases are re-read with AI assistance (Fig. 4B, C). Key features: (1) Washout. To minimize memory effects, a sufficiently long washout period must be implemented between sessions; its duration depends on the number and characteristics of cases[76] and is typically at least 4 weeks[74,89]. (2) Reshuffle. The case order must be reshuffled in the second session to further reduce memory and sequencing effects. (3) Risk of learning effects. Readers are generally more familiar with the interface and case structure during the second session, which may artificially improve performance and be misattributed to AI assistance.

3. Crossover design. It addresses the sequencing effects of sequential-session designs and is the most recommended paired-study structure. Readers are randomly assigned to two sequences: sequence A completes unassisted-AI followed by AI-assisted after a washout period, whereas sequence B follows the reverse order, forming a 2 × 2 crossover (Fig. 4D). When multiple assistance modes are included, a 3×3 crossover can be implemented (Fig. 4E). Key features[76,92,173]: (1) Elimination of sequencing effects. By balancing reading sequences, the crossover design effectively controls for order-related factors such as interface familiarity and fatigue. (2) Self-controlled structure. It preserves the statistical efficiency and low sample-size requirements of paired designs while incorporating the methodological rigor of randomized controlled trials, making it the most robust paired-design format. (3) Washout period. As in sequential-session designs.

4. Variants of paired designs. The above represent standard paired-design structures, but practical constraints often necessitate extensions that balance workload, order-effect control, and case allocation (Fig. 4F–H). Split-sample sequential designs with two or three case subsets remain fundamentally paired: each clinician re-reads the same case group under different conditions, with case splitting and reader grouping used to reduce individual workload (Fig. 4F–G). In contrast, the three-way crossover design adopts a Latin-square structure but assigns different case sets to each condition[67,83]. Because clinicians do not re-read the same cases, it is no longer a paired design but a non-paired multi-sequence crossover (Fig. 4H).

Key design elements. The exploratory paired design relies on a rigorous self-controlled structure and minimizes memory and order effects[76,77]. A representative dataset must first be assembled, with reference standards established by an independent expert panel or pathology. To mitigate memory and order effects, a washout period is typically implemented, supplemented by the random reshuffling of cases in the second session[76,77]. Besides, the reading interface should remain identical across sessions, with AI outputs (e.g., probability scores, heatmaps, or classification labels) introduced only in the second session. For each reading, clinicians' final decisions, confidence ratings, reading time, and degree of reliance on AI should be recorded.

To rigorously isolate true automation bias from confounding interaction variables, these studies must employ strict experimental controls. While the inclusion of trap cases (intentionally incorrect AI predictions) is helpful for probing over-reliance[78], it is not sufficient on its own. Protocols must explicitly standardize or systematically vary interface elements, such as alert saliency, heatmap quality, and the presentation of predictive confidence, to ensure that UI-driven coercion is not conflated with genuine cognitive bias[79]. Furthermore, in addition to random reshuffling to mitigate case-order and fatigue effects, studies should mandate a dedicated familiarization phase (a warm-up set of cases) prior to actual testing[80]. This effectively separates the clinician's initial learning curve from their steady-state cognitive coupling with the AI.

The sample sizes for the three paired designs are estimated based on the specific metric types, such as mean differences, paired proportions, or the covariance structure of correlated AUCs. Accordingly, the main distinction among the designs lies not in the formula itself but in the estimation of the correlation coefficient (ρ)[81]: crossover designs, assisted with washout periods, better control memory effects and therefore yield correlation estimates closer to real-world diagnostic agreement. In contrast, case-by-case sequential designs typically exhibit the highest correlation due to memory effects, resulting in the smallest theoretical sample size but also the greatest risk of bias[76,77,82]. Although statistical power is driven largely by case volume, at least three clinicians of differing seniority should be included as repeated-measurement sources to account for reader variability (Supplementary Material 4).

Evaluation components. The evaluation framework centers on changes in diagnostic performance (Δ performance), reading efficiency (Δ reading time)[67,83], and may incorporate exploratory clinical utility via DCA when probability-based outputs are available[41]. It also evaluates how clinicians interpret and respond to AI uncertainty displays (e.g., confidence intervals or warning prompts), including error-correction rates and the proportions of false-to-true or true-to-false revisions, to assess their effects on reliance behavior, automation bias, diagnostic confidence, and potential risks of over- or mis-diagnosis[84,85]. In addition, changes in diagnostic confidence (Δ confidence), human–AI confidence calibration, and intra-reader agreement provide deeper insight into clinicians' cognitive and decision stability during AI-assisted interpretation[67,83].

Strengths and limitations. The design offers high sensitivity and efficiency[76,77], enabling the detection of subtle AI-induced cognitive effects, including error correction, changes in confidence, shifts in reading time, and automation bias, through its self-controlled structure[77]. Conducted in a controlled, risk-free environment, it provides a low-cost signal-detection step, generating directionally informative evidence for interface optimization and subsequent design of higher-level validation studies. Its limitations arise from the fact that findings primarily reflect short-term cognitive changes under controlled conditions rather than effects on real clinical workflows or patient outcomes, limiting generalizability. And it cannot eliminate memory effects, nor can it assess the engineering stability of AI under authentic data streams.

Method 5. Multi-reader multi-case (MRMC)

Rationale and positioning. MRMC studies are among the most confirmatory designs for evaluating interpretive diagnostic AI systems. They aim to assess, under controlled conditions, the impact of AI assistance on clinical diagnostic performance while quantifying the dual sources of variability arising from readers and cases. Unlike exploratory paired designs, MRMC provides statistically powered confirmatory evidence to address two core questions: (1) Does AI significantly improve clinicians' diagnostic performance? and (2) Is this improvement consistent across different readers and cases? MRMC has been explicitly endorsed by the FDA and EMA for evaluating AI-assisted diagnostic devices and serves as a key evidentiary framework for demonstrating human–AI performance gains in regulatory submissions for high-risk medical software[86,87].

Overview of MRMC methodology. MRMC requires multiple clinicians (typically of varying seniority, with at least five readers[88,89]) to diagnose the same case set under different reading modes[86]. The statistical foundation of MRMC analysis rests upon a dual-random-effects structure encompassing both "Reader × Case". In the context of AI evaluation, the inclusion of different reading modes (e.g., unassisted vs AI-assisted) generates a three-dimensional data structure: "Reader × Case × Mode"[86,90]. There are three designs of MRMC, summarized in Table 1[86,89,91].

Table 1 | Comparative analysis of three MRMC study designs

Characteristics	Fully crossed design (Paired readers and paired cases)	Partially crossed design I (Unpaired readers and paired cases)	Partially crossed design II (Paired readers and unpaired cases)
Core architecture	All readers evaluate all cases using all reading modes.	Different groups of readers evaluate different modes, but all groups evaluate the same set of cases.	The same group of readers evaluates all modes, but each mode is evaluated on a different, independent set of cases.
Statistical capacity	Controls for both inter-case and inter-reader variability; allows estimation of reader-case interactions.	Controls for inter-case variability. Effectively eliminating confounding from case heterogeneity.	Controls for inter-reader variability. Effectively eliminating confounding from varying reader expertise levels.
Potential bias	Learning effect; Memory effects.	Inter-reader variability: Differences in seniority between reader groups may confound mode comparisons.	Inter-case variability: Differences in disease spectrum or difficulty between case sets may confound mode comparisons.
Estimable interactions	All interactions: case-mode, reader-mode, reader-case, and higher-order interactions.	Only allows estimation of case-mode interaction	Only allows estimation of reader-mode interaction
Feasibility	Low. Extreme workload for readers; prone to fatigue; challenging for large-scale implementation.	High. Readers only need to master one mode, halving the workload and simplifying organization. Requires rigorous random assignment of readers.	Moderate. Readers need to master two modes, halving the workload. The primary challenge is ensuring high comparability between the two case sets.

MRMC multi-reader multi-case.

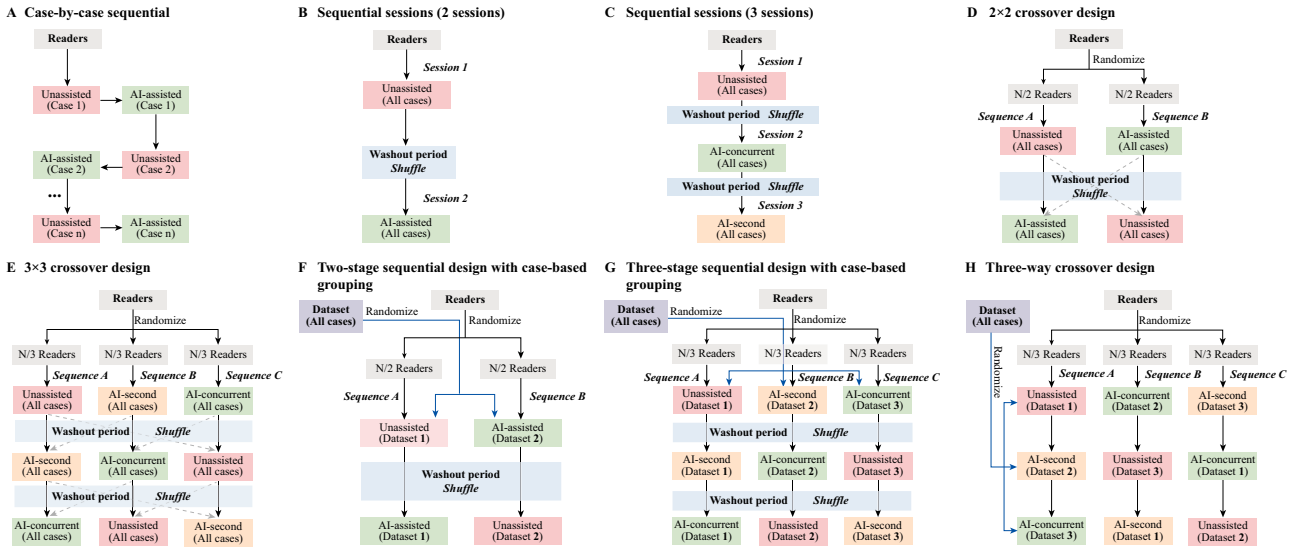


Fig. 3 | Schematic overview of paired designs and their variants. A Case-by-case sequential; **B** Two-stage sequential sessions; **C** Three-stage sequential sessions; **D** 2 × 2 crossover design; **E** 3 × 3 crossover design; **F** Two-stage sequential design with case-based grouping; **G** Three-stage sequential design with case-based grouping; **H** Three-way crossover design. Note: AI-second and AI-concurrent represent two distinct modes of AI assistance[83]. AI artificial intelligence.

The fully crossed MRMC design[86,91] is the most strongly recommended structure. All clinicians diagnose the entire case set under two or more reading modes, generating a complete "N(cases) × n(readers) × M(modes)" three-dimensional data matrix (Fig. 3A–C). This configuration provides the highest statistical efficiency and the most stable estimation of mode effects, albeit at the cost of substantial reader workload. Although the schematic logic of Fig. 3B, C (MRMC designs) resembles that of Fig. 4D, E (crossover designs), they diverge fundamentally in terms of variable dimensionality, sources of error, sample size estimation, and statistical inference.

Partially crossed (split-plot) MRMC design[91]. To reduce reader burden, partially crossed structures may be used: (1) Unpaired readers, paired cases: all cases appear in both modes, but readers are divided into groups, each assigned to only one mode (Fig. 3D–F). (2) Paired readers, unpaired cases: all readers interpret both modes, but the case set is split into subsets, with each case read in only one mode (Fig. 3G–I). These designs substantially reduce per-reader workload and allow inclusion of larger case sets, but they introduce structural missingness into the data matrix. As a result, variance estimation and model fitting become more complex, and statistical efficiency is lower than in fully crossed MRMC.

Relation to crossover designs. Although MRMC studies often employ crossover procedures, the two are conceptually distinct and operate at different inferential levels. The purpose of the crossover design is to control learning and order effects, ensuring comparability of a clinician's performance across reading modes. Its inference is confined to the "reader-case" self-controlled unit, addressing whether AI changes an individual clinician's diagnosis of a specific case[76,92]. In contrast, MRMC focuses on population-level inference. By incorporating multiple readers and a large case set, it constructs a three-dimensional "reader × case × mode" data structure that enables estimation of the average AI effect across reader and case populations, as well as the stability of that effect[88]. Although both approaches involve multi-session reading, randomized sequence, and washout periods, MRMC operates at a higher inferential tier and provides stronger external validity. Detailed comparison is shown in Supplementary Table 1.

In high-quality study practice, crossover designs and MRMC are complementary. Crossover procedures provide a bias-controlled operational framework, whereas fully crossed MRMC supplies the data structure required for confirmatory inference. The former addresses "how to arrange reading sequences", while the latter answers "whether the AI effect generalizes across readers and cases". Accordingly, MRMC is typically implemented using paired designs with sequence randomization to ensure identifiability of mode effects and to avoid temporal confounding (Fig. 3B, C, H, I)[89]. If a fixed reading order is used (Fig. 3A, D–G), mode effects may become confounded with time effects, preventing accurate attribution of the AI's true contribution.

Key design elements. The design elements of MRMC studies largely mirror those of paired designs: a representative dataset, an independently established reference standard, and readers with varying levels of seniority[89]. Reading procedures must adhere strictly to blinding and randomization principles, with adequate washout periods between modes and reshuffling[89]. The reading interface should remain identical across modes, with AI outputs added only in the assisted condition. When appropriate, balanced or Latin-square arrangements may be used to further mitigate learning effects.

MRMC analysis relies on specialized dual-random-effects frameworks rather than conventional difference testing. The Obuchowski-Rockette (OR) and Dorfman–Berbaum–Metz (DBM) models explicitly account for the covariance structure by decomposing reader effects, case effects, and reader-by-case interactions, thereby avoiding underestimation of standard errors and yielding robust effect-size estimates[88]. This modeling framework is a defining methodological feature that distinguishes MRMC from external evaluation, shadow-mode, and paired designs, and is a key reason for its broad acceptance by regulatory agencies.

Current mainstream MRMC analysis tools include the OR-DBM MRMC software[93–95] suite and the R package MRMCaov[96]. In parallel, iMRMC, developed and maintained by the US FDA, has become a standard tool in this field[87]. iMRMC offers greater flexibility for handling unbalanced designs, where not all readers evaluate all cases, and can automatically accommodate missing data. For studies intended to support FDA submissions, for complex sample-size simulations, or for datasets with substantial missingness, iMRMC is generally the preferred option.

Sample-size determination in MRMC requires simultaneous modeling of reader- and case-level random effects and their cross-correlation structure. As a result, traditional analytic formulas for diagnostic studies are not applicable. Instead, the required numbers of cases and readers must be derived from effect-size assumptions and variance components under the OR/DBM or U-statistic frameworks. When pilot data are available, the

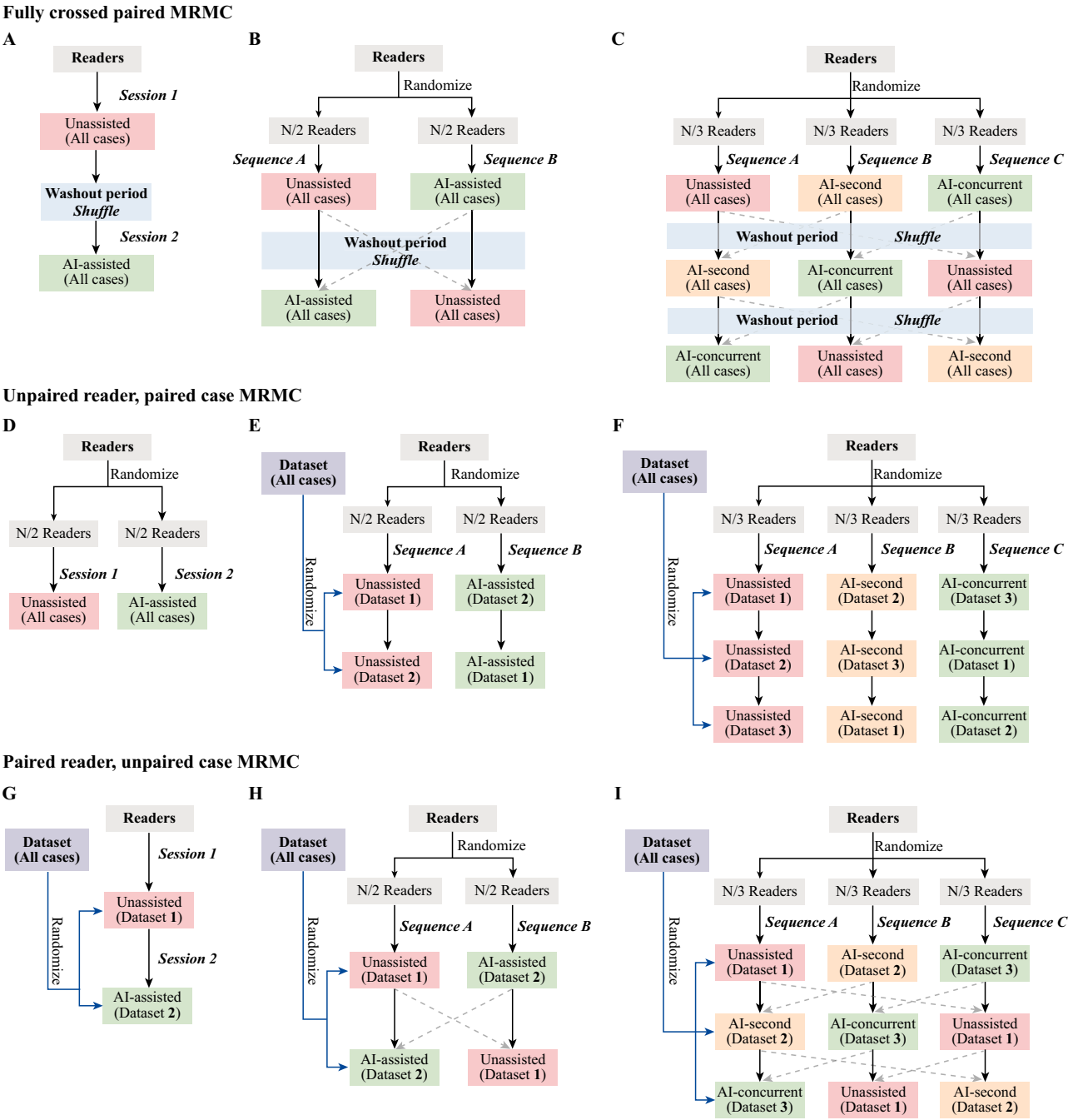


Fig. 4 | Schematic overview of multi-reader multi-case. A–C Fully crossed paired MRMC. **A** Fixed-sequence fully crossed MRMC; **B** Two-sequence randomized fully crossed MRMC; **C** Three-sequence randomized fully crossed MRMC. **D–F** Unpaired reader, paired case MRMC. **D** Fixed-sequence unpaired reader, paired case MRMC; **E** Two-sequence randomized unpaired reader, paired case MRMC; **F** Three-sequence randomized unpaired reader, paired case MRMC. **G–I** Paired reader, unpaired case MRMC. **G** Fixed-sequence paired reader, unpaired case MRMC; **H** Two-sequence randomized crossed paired reader, unpaired case MRMC; **I** Three-sequence randomized crossed paired reader, unpaired case MRMC. Note: AI-second and AI-concurrent represent two distinct modes of AI assistance[83]. AI artificial intelligence, MRMC multi-reader multi-case.

OR-DBM MRMC software can compute sample size using empirically estimated variance components[91,96–98]. In the absence of pilot data, MRMC sample size[88,99] or iMRMC[87] can be used to obtain approximate case-reader combinations through assumed correlation structures or simulation (Supplementary Material 5).

Evaluation components. The MRMC evaluation framework typically spans five components: primary effect metrics (e.g., ΔAUC), variance decomposition for reader and case variability, statistical inference using DBM/OR F-tests, reader- and case-level analyzes, and assessments of agreement and bias[87,100,101]. In applied medical studies, the most commonly reported metrics include the reader-averaged AUC, the AUC difference (ΔAUC), sensitivity and specificity, and significance testing based on DBM/OR models[28,29,102]. Many studies additionally report stratified reader analyzes to examine whether AI provides differential benefit across experience levels[29].

Strengths and limitations. MRMC offers strong confirmatory value, statistical rigor, and generalizability[94,103]. It enables simultaneous estimation of the average AI's performance gain, inter-reader variability, and

case-difficulty variability, providing more robust evidence than single-reader or paired designs[88,89,104]. In addition, MRMC also supports stratified analyzes to characterize differential AI benefit across reader experience levels or case-complexity strata. However, MRMC studies are resource-intensive, requiring substantial reader recruitment, and imposing a considerable reading burden, and they cannot fully replicate real clinical workflows. Accordingly, MRMC should be viewed as a key confirmatory step bridging cognitive validation and clinical validation, providing high-quality performance evidence and behavioral baselines to inform subsequent prospective clinical trials.

Phase 4. Clinical evidence validation: confirming clinical benefit and patient outcomes

Rationale and positioning

The clinical evidence phase marks a fundamental shift in evaluation logic, from assessing "model performance" to determining "interventional benefit". Its core mission is to evaluate whether a medical-AI system, once embedded into real clinical workflows, can optimize clinicians' decision pathways and ultimately translate into measurable clinical gains, such as improved diagnostic efficiency, reduced unnecessary interventions, or better long-term patient outcomes[105–107]. This phase represents the critical juncture at which medical AI must overcome the "performance wall" and generate evidence suitable for clinical adoption; the resulting data directly influence regulatory approval and guideline endorsement[108,109].

This phase typically employs a multidimensional framework based on prospective randomized controlled trials (Method 6), which use controlled comparisons to eliminate confounding and ensure that any observed benefit can be causally attributed to AI[110,111]. Depending on the validation objective, three major trial designs are commonly used[110] (Box 2): diagnostic randomized controlled trials (RCT), clinical-effectiveness RCT, and stepped-wedge cluster randomized trials (SW-CRT), a specific form of cluster randomized controlled trial. Together, these designs form the core methodological system of the clinical evidence phase, addressing AI's value from complementary perspectives of diagnostic behavior, patient-level benefit, and system-level deployment. A detailed comparison is provided in Supplementary Table 2. Furthermore, to ensure absolute accountability and high-fidelity evidence generation, prospective interventional trials in Phase 4 must strictly utilize the SPIRIT-AI extension[112] during the clinical trial protocol design stage and adhere to the CONSORT-AI extension[113] when reporting final trial outcomes.

Key design elements. Crucially, prior to trial initiation, a formal study protocol and SAP must be explicitly prespecified. In this phase, prospective RCTs provide the most compelling evidence for establishing a causal association between AI intervention and clinical benefit[108,114]. These trials adhere to the principles of "randomization, control, intervention, and prospective", thereby minimizing confounding and ensuring baseline comparability between study arms[110]. At the design stage, investigators must specify the AI system's functional role within the clinical pathway, such as primary screening, second-opinion support, prioritization, or workflow triage[67,115], and select an appropriate trial design accordingly (Box 2).

Regardless of the specific design, several requirements are essential for this phase: the AI interface, thresholds, and alerting logic must remain fixed throughout the trial; participants should be enrolled consecutively or through natural exposure to avoid selective inclusion of atypical cases; AI outputs, clinician decision pathways, reading time, coverage, and failure rates should be systematically recorded; and safety-monitoring mechanisms (e.g., predefined stopping rules) should be established for high-risk scenarios. In multicenter studies, differences in workflow, equipment, and staffing must be documented and controlled to enhance generalizability. While each of the three designs has its own specific evaluation metrics (as detailed in Box 2), Phase 4 trials evaluating probabilistic risk scores or clinical decision-support systems should explicitly incorporate DCA to quantify clinical net benefit[41], alongside rigorous assessments of interventional calibration and uncertainty to ensure safe human–AI decision-making. Furthermore, to align with the real-world mandates of health technology assessment bodies, these trial designs can be methodologically extended to incorporate health-economic endpoints[116,117]. By concurrently collecting cost and health-utility data (e.g., QALYs) alongside primary clinical efficacy outcomes, this phase provides a rigorous foundation for trial-based cost-effectiveness analysis, generating the incremental cost-effectiveness ratios necessary for reimbursement.

To address structural sources of site heterogeneity, this phase must incorporate design and analytic strategies that go beyond simple sample-size inflation. First, AI-by-site interaction testing is required to determine whether the interventional effect of AI differs systematically across centers with distinct case-mix profiles, acquisition-process characteristics, or clinical workflows. Second, for probability-output models, guidance for local threshold recalibration should be prespecified to account for center-level prevalence shifts and case-mix variation[118,119], ensuring that decision thresholds remain clinically appropriate across heterogeneous settings. Third, statistical analysis strategies can rigorously distinguish the direct effect of the AI from site-

Box 2. | Common types of prospective randomized controlled trials

1. Diagnostic RCT: This design evaluates whether AI improves diagnostic behavior or workflow efficiency within real clinical settings. Patients or cases are randomized to AI-assisted versus standard-of-care arms, and clinicians use or do not use AI according to their assigned workflow. Investigators compare surrogate endpoints, such as diagnostic performance, reading time, and workflow latency, to quantify AI's direct impact on diagnostic results and process efficiency[26,174]. As the most fundamental and widely used prospective RCT design for medical AI, it is well-suited to triage and screening settings and is effective in capturing behavioral effects under real-world workload and case complexity, although it does not directly evaluate patient-level clinical outcomes.

2. Clinical effectiveness RCT: This design parallels the diagnostic RCT but shifts the focal point from "diagnostic performance" to tangible "patient benefit', representing the pinnacle of the evidence hierarchy for medical AI. Common endpoints include mortality, complication rates, readmission, length of stay, treatment modification, and key time-sensitive process metrics (e.g., door-to-needle time)[106,107]. Given the lower incidence of clinical events, such trials typically necessitate larger sample sizes and extended follow-up periods to ensure adequate statistical power. This evidence level is pivotal for the integration of AI into clinical guidelines and its adoption into healthcare policy.

3. SW-CRT: This design is a cluster-level design in which clinical units (e.g., departments, hospitals, or communities) are randomized to transition from the control to the intervention condition at prespecified time points[114,175]. This structure reflects real-world AI deployment across multiple centers, reduces spillover effects arising from within-institution exposure to AI, and is ethically favorable because all clusters eventually receive the intervention[114,176]. SW-CRT enables evaluation of AI's organizational-level impact with high real-world implementation validity. However, its implementation requires careful handling of time trends, intra-cluster correlation, and delayed intervention effects to ensure accurate estimation of the AI-specific effect[177,178].

RCT, randomized controlled trial; SW-CRT, stepped-wedge cluster randomized controlled trial

specific practice effects by employing hierarchical or mixed-effects models, site-stratified analyses, or cluster-robust inference[120,121]. These measures align with CONSORT-Cluster[122] and TRIPOD-Cluster[119] guidelines, ensuring that this phase evaluates genuine causal clinical benefit rather than artifacts driven by center-specific practice patterns.

Sample size calculation. Sample-size determination in RCTs depends primarily on the hypothesis-testing framework[123,124]. In clinical evidence studies of medical AI, superiority and non-inferiority designs are most commonly used, whereas equivalence designs are rarely adopted[28,105,125,126]. This pattern reflects the functional role of AI systems, their intended clinical value, and regulatory expectations. Superiority trials assess whether AI can meaningfully outperform standard care on key metrics such as sensitivity, missed-diagnosis rate, or workflow efficiency, making them suitable for systems aimed at performance enhancement[124]. Non-inferiority trials emphasize safety and substitutability, evaluating whether AI can deliver advantages in speed, cost, or accessibility without materially compromising accuracy[127,128], an approach particularly relevant to triage, screening, and resource-limited settings.

Equivalence designs are uncommon in medical AI because they require two-sided margins and substantially larger sample sizes than superiority or non-inferiority trials[72,124,125]. Moreover, AI systems are typically intended to improve, or at least not fall below, clinician performance rather than achieve strict equivalence within a predefined interval; behavioral differences between clinicians and AI further challenge the plausibility of equivalence assumptions. Regulatory agencies also focus on demonstrating effectiveness and safety rather than statistical equivalence to human diagnosis[129]. Consequently, superiority and non-inferiority frameworks constitute the primary inferential basis for medical-AI clinical trials and underpin subsequent decisions on sample-size estimation, effect-size specification, and interpretation of results. The three trial types require distinct sample-size calculation frameworks, respectively, see Supplementary Material 6 for details.

Strengths and limitations. The phase offers the advantage of evaluating the real-world effectiveness of AI within real clinical workflows, providing evidence that most closely reflects routine practice[26,108,110]. However, these trials are resource-intensive, lengthy, and susceptible to variability in workflow, clinician behavior, learning effects, and inter-center heterogeneity[130,131]. Blinding is also difficult to maintain in real clinical environments, introducing potential behavioral bias. Despite these challenges, this phase remains essential for moving from performance validation to clinical value validation and provides the definitive evidence required to support clinical deployment of AI systems.

Phase 5. Real-world integration validation: post-market performance and system-level impact

Rationale and positioning. The real-world integration phase focuses on the long-term performance of medical-AI systems within real healthcare ecosystems. Its core question shifts from isolated "interventional efficacy" to the robustness, fairness, and organizational assimilation of AI across institutions, populations, and periods. By examining the dynamic interactions between AI and hospital structures, clinical workflows, clinician skill evolution, and regulatory systems, this phase aims to characterize the long-term effects and systemic adaptability of AI within complex health-care ecosystems. Phase 5 focuses on real-world clinical deployment, system-level integration, and post-market performance governance. Unlike Phase 2, which functions as a non-interventional silent trial to evaluate engineering stability under real-world data streams without influencing clinical decisions, Phase 5 assesses the AI system's sustained clinical impact, equity, safety, adoption, and long-term performance once it is actively integrated into live, routine care.

Compared with phase 4, the core advantage of the phase lies in its ability to evaluate the system-level resilience of AI under real-world noise and operational complexity[132]. To safeguard internal validity, controlled studies such as MRMC and RCTs often exclude "noisy data", including poor-quality images, missing information, or workflow anomalies[28,29]. However, such data are ubiquitous in routine clinical practice. Therefore, by incorporating real-world data noise, workflow perturbations, and inter-institutional heterogeneity, this phase assesses both the technical robustness of AI under non-ideal conditions and its real-world adaptability through implementation-science metrics, such as adoption, fidelity, and fairness, thereby addressing the limitations in external validity inherent to clinical trials.

Importantly, the monitoring function of the real-world integration phase does not supersede RCTs, MRMC studies, or other controlled validation designs[133,134]. The objective of this phase is intended to extend rather than replace clinical-trial evidence: its purpose is to ensure that AI remains effective in routine practice, not to establish its initial "efficacy". Accordingly, the real-world integration and clinical evidence phases maintain a complementary and synergistic relationship within the medical AI evaluation hierarchy, jointly supporting safety and effectiveness across the full lifecycle of AI systems.

The core methodological approach in this phase is prospective real-world implementation and integrated evaluation (Method 7), which encompasses four complementary study types: single-arm prospective pre-post study, real-world effectiveness study (RWS), implementation research (IR), and post-market active surveillance (See Box 3 for details). Together, these approaches form the post-market evidence chain for medical AI and are essential for ensuring that AI systems are "usable, useful, and sustainably effective" in real-world settings.

Key design elements. Crucially, prior to real-world implementation, a formal monitoring protocol or a versioned analysis charter must be explicitly prespecified. These approaches typically use observational or mixed-methods designs and follow a pragmatic-research paradigm: they prospectively track consecutively enrolled real-world populations without altering clinical decision-making, thereby assessing the actual utility and safety of AI in routine practice. Appropriate methodological choices should be made according to study objectives (See Box 3 for details). Regardless of the specific approach, the phase must ensure that AI versions, thresholds, and interfaces remain stable and traceable throughout the study; that failure modes and anomalous events arising in real workflows are systematically documented; that performance consistency is monitored across institutions, devices, and patient groups; and that comprehensive records of data governance, privacy protection, and model-update processes are maintained. Each of the four designs has its own evaluation metrics (Box 3). Furthermore, to facilitate institutional scaling and long-term sustainability, these real-world deployment designs can be methodologically augmented with health-economic decision analytics. This includes budget impact analysis to project the net financial footprint of AI adoption on local healthcare budgets over a multi-year horizon[135], and value-of-information analysis[136] to mathematically assess whether the reduction of evidence uncertainty justifies the economic costs of delaying broader clinical integration. Additionally, to support the global mandate for green healthcare, longitudinal monitoring in Phase 5 must track the cumulative compute costs and the long-term operational carbon footprint associated with continuous algorithmic maintenance and periodic retraining[137].

Sample size calculation. In terms of sample size, this phase moves away from traditional hypothesis-testing frameworks and instead emphasizes precision-driven and risk-coverage principles. First, real-world studies typically do not specify a strict H_0 but aim for adequate "coverage", ensuring that estimates achieve a prespecified level of precision. Second, when monitoring rare adverse events (e.g., severe misdiagnosis), required sample sizes are derived using the "rule of three" or binomial-based back-calculation[55–57]. Finally, for common performance indicators such as misdiagnosis, confidence-interval width methods are commonly applied[52–54].

Strengths and limitations. The core value of the phase lies in its high real-world representativeness. Compared with tightly controlled RCTs,

Box 3. | Common types of prospective real-world implementation and integrated evaluation

Common types

1. Single-arm prospective pre-post study. This design compares clinical indicators before and after AI implementation within the same department or institution, using the introduction of AI as the temporal intervention point. It primarily evaluates changes in workflow efficiency, diagnostic pathways, clinician behavior, and resource utilization following deployment[179,180]. Rigorous attention must be paid to maintaining comparability in case mix and resource conditions across the pre- and post-AI periods. Detailed documentation of AI coverage and clinicians' reliance on or disregard of the system is essential for quantifying workflow gains attributable to the intervention.

2. Real-world effectiveness study (RWS). RWS evaluates the external validity and system-level robustness of AI using uncontrolled data sources such as electronic health records and regional clinical data repositories[138,181]. By retaining "real-world noise", including device variability and workflow heterogeneity, this design examines how AI performs when exposed to diverse populations, atypical cases, and clinicians with varying levels of experience. It thereby reveals the "efficacy-effectiveness gap" that emerges when AI transitions from idealized experimental settings to routine clinical practice.

3. Implementation research (IR). Using implementation-science frameworks, IR extends beyond clinical endpoints to evaluate non-clinical determinants of successful adoption, including acceptability, fidelity, and sustainability[182,183]. Its central aim is to address the persistent gap between technical efficacy and real-world uptake. Through mixed-methods approaches that combine quantitative indicators with qualitative interviews, IR identifies organizational barriers and clinician-level behavioral drivers that hinder large-scale deployment of AI, thereby providing methodological guidance for optimizing human–AI collaboration.

4. Post-market active surveillance. This approach establishes automated monitoring pipelines for longitudinal safety, equity, and performance tracking[184,185]. Operationalizing post-market surveillance requires a risk-stratified monitoring cadence, with high-risk diagnostic systems undergoing daily or weekly automated checks and lower-risk applications monitored on a monthly basis. Statistical process-control methods (e.g., CUSUM, EWMA, and Shewhart control charts) should be deployed to dynamically detect gradual or abrupt drift in real-world data streams. When prespecified alerting thresholds are breached, such as ≥5% AUROC decline, ≥10% sensitivity loss, subgroup-specific degradation, fairness-drift thresholds, or safety-critical error triggers, a structured de-implementation pathway must be automatically activated. This pathway includes clear fall-back protocols, such as automatic rollback to a prior validated version, reverting the model to background shadow-mode tracking, temporary suspension of deployment, or full withdrawal, thereby safeguarding patient care while root-cause analysis and model retraining occur.

Interconnectivity of methodologies

Together, these four approaches form a sequential and complementary evidence chain. Single-arm pre-post studies detect immediate workflow changes at the point of deployment. Real-world effectiveness studies then evaluate pragmatic impact and external validity under minimally controlled conditions. Implementation research explains the organizational and behavioral determinants of successful adoption and sustained use. Finally, post-market active surveillance provides continuous monitoring of long-term safety, fairness, and performance drift across sites and populations. Collectively, they establish a closed-loop pathway spanning deployment, effectiveness, implementation, and long-term safety, enabling scalable and trustworthy clinical adoption of medical AI.

studies in this phase typically involve larger sample sizes and broader population representativeness, enabling the detection of rare risks that may be missed due to limited sample size or follow-up duration in trials, and providing evidence with strong external validity[138,139]. Its major limitation is the absence of randomization, which increases susceptibility to confounding and weakens causal inference relative to RCTs[133]. In addition, real-world data vary in quality and often contain missing information and multiple sources of bias[138,140]. Pre-post designs are vulnerable to time-trend bias and concurrent policy changes; RWS may not fully correct for selection bias arising from non-random exposure; IR relies on qualitative interpretations; and active surveillance requires robust data infrastructure and sustained resource investment. Moreover, the effects of AI may fluctuate across time, institutions, and patient groups in complex real-world environments.

Core driver. Dynamic evaluation architecture: a meta-methodology for evolving AI

Rationale and positioning. The defining distinction between medical AI and traditional medical devices lies in their rapid iterability. Medical AI functions as an "evolving system", with performance that fluctuates in response to shifts in data distribution, device characteristics, and clinical workflows, and with model versions updated far more frequently than the timelines of conventional clinical trials[141]. The traditional linear paradigm of "development-validation-deployment" is therefore insufficient, creating structural challenges of validation lag and evidence obsolescence[142,143]. This necessitates a meta-methodological framework capable of spanning the entire lifecycle and adapting dynamically as models evolve.

The dynamic evaluation architecture provides such a lifecycle-wide meta-logic and is theoretically applicable across all stages of medical-AI evaluation, particularly during the phases 4 and 5. However, current literature and regulatory practice indicate that adaptive designs remain largely conceptual and methodological, with limited mature application in medical AI-assisted systems[141]. Implementing adaptive designs in early phases may compromise reproducibility and stability of performance assessment, whereas their use in clinical trials faces ethical, regulatory, and infrastructural constraints[144].

The US FDA's Predetermined Change Control Plans (PCCPs) provide a practical regulatory pathway to address this challenge[145]. PCCPs require manufacturers to pre-specify the scope of allowable modifications, the associated validation procedures, and benefit-risk mitigation strategies at the time of initial submission, with prospective authorization granted upon approval. This mechanism enables rapid algorithmic iteration within a regulator-approved boundary, without the need for repeated marketing submissions. As the most explicit regulatory mechanism to date, PCCPs transition dynamic evaluation from theoretical frameworks into institutionalized practice, establishing a robust cornerstone for the continuous learning and iterative optimization of medical AI. Crucially, this architecture resolves the paradox between the strict frozen-model requirement of early phases and the reality of continually learning models. Once a safety baseline is established (Phases 1–3), models that continuously learn from real-world data streams can utilize mechanisms like PCCPs to dynamically update within pre-approved boundaries, circumventing the need to restart the entire evaluation pipeline from Phase 1.

Box 4. | Common types of adaptive design

1. Adaptive sample size re-estimation (SSR). SSR is a well-established adaptive design technique that dynamically adjusts the total sample size based on prespecified interim analyzes to address uncertainty in the assumed effect size or variability at the design stage[186,187]. In medical-AI trials, SSR allows flexible sample-size modification when the anticipated performance gain of the AI system is underestimated, thereby preserving statistical power and reducing the risk of type II error. However, implementing SSR in heterogeneous real-world studies, where data-quality issues and multiple confounders are common, poses substantially greater challenges for bias control and statistical robustness, and entails more complex regulatory considerations than in traditional controlled experiments.

2. Adaptive platform trials (APT). APTs use a master protocol to integrate multiple sub-studies, allowing different interventions, such as multiple AI models, to enter or exit dynamically while sharing control groups and trial infrastructure[147,188,189]. This design aligns closely with the parallel development and rapid iteration characteristic of medical AI, substantially improving evaluation efficiency and avoiding the need to launch separate trials for each new model or version. APTs are therefore regarded as an ideal infrastructure for building a continuous-evaluation ecosystem for medical AI. However, successful implementation requires substantial upfront investment, including harmonized data standards, centralized review processes, and sophisticated statistical support.

3. Versioned evaluation (VE). As a dynamic methodology tailored for medical AI, VE treats algorithmic updates (e.g., v1 to v2) as a "dynamic intervention", allowing for intra-trial upgrades while maintaining independent, comparable assessments for each version[145]. Unlike traditional RCTs that mandate a "fixed intervention", this approach accommodates the algorithmic fluidity necessitated by data drift or hardware transitions, preventing trial results from becoming obsolete before completion[149]. By establishing version-tracking mechanisms, transition protocols, and drift monitoring systems, it ensures methodological rigor throughout the AI's evolutionary cycle[149]. This framework effectively reconciles the discrepancy between rapid AI iteration and protracted clinical trial timelines, serving as a cornerstone for evaluating dynamic deployment in medical AI.

Key design elements. Adaptive design (Method 8) is a key methodological tool enabling medical-AI diagnostic systems to achieve continuous evaluation, updating, and safety. Among the various adaptive approaches, three designs currently offer the greatest feasibility and methodological value: adaptive sample size re-estimation (SSR), adaptive platform trials (APT), and versioned evaluation (VE). Together, these designs represent the most suitable adaptive-design strategies for contemporary medical-AI evaluation (Box 4).

Currently, sample size planning for adaptive designs in medical AI has shifted from traditional fixed-hypothesis testing to rule-based adaptive strategies. Rather than a static figure, the sample size is driven by the "informational threshold" required for robust inference (e.g., Bayesian posterior probability) and the "monitoring breadth" necessary to capture heterogeneous risks[146]. Because interim adaptations are permitted, sample-size determination typically relies on statistical simulation rather than a single closed-form formula[146–148]. Consequently, Regulatory review focuses on the scientific validity and transparency of the adaptive mechanisms, including pre-specified triggering conditions, rigorous control of type I error, and operational integrity to ensure unbiased implementation[146,148]. Furthermore, this dynamic meta-logic allows for the selective implementation of the framework based on the clinical task. For example, non-diagnostic AI applications, such as operational algorithms for bed allocation or administrative triage, typically do not require the rigorous cognitive human–AI interaction assessments of Phase 3. Instead, they can selectively abbreviate the pathway, leveraging APT within Phase 5 to continuously evaluate system-level efficiency and workflow impact.

Evaluation components. The evaluation metrics for adaptive designs are fundamentally determined by the validation phase in which they are applied (e.g., RCTs or real-world integration phase), with additional emphasis on temporal stability and model evolution[27,145]. Key incremental metrics include the degradation rate, quantifying the decline in AI performance over time; calibration drift, capturing the divergence between predicted probabilities and observed outcomes; and version-transition shock, assessing the transient disruption to clinical workflows or diagnostic outputs following algorithm updates. This synthesis of phase-specific and evolutionary metrics ensures precise oversight of AI performance dynamics throughout its lifecycle.

Operationalizing lifecycle management. To ensure the safe and scalable deployment of dynamic AI architectures, robust lifecycle-management rules must be operationalized. However, because clinical risks vary drastically across different AI tasks, establishing universal, rigid quantitative thresholds for performance drift or system rollbacks is neither feasible nor scientifically sound[141]. Instead, drawing upon emerging machine learning operations (MLOps) frameworks for clinical quality improvement[149], we propose a structured clinical lifecycle-management matrix (Table 2) that operationalizes dynamic evaluation through predefined action and transition rules. Crucially, while static testing dominates the early stages of AI development, this dynamic MLOps loop is predominantly enacted during Phase 4 (interventional trials) and Phase 5 (real-world integration), where models must continuously adapt to real-world interactions and shifting clinical environments[141].

Drawing upon regulatory paradigms such as the FDA's PCCPs[145] and algorithmic vigilance strategies for mitigating dataset shift[150], this matrix delineates the core guardrails across an iterative, five-step loop (Step A to E): from pre-deployment PCCP setup to active clinical surveillance, trigger evaluation, adaptation, and post-update governance. It outlines illustrative, task-dependent clinical metrics, such as case-mix shifts, ECE, and human–AI behavioral drift, and explicitly defines the gating conditions under which an AI system should trigger local recalibration, execute an immediate rollback to prevent version-transition shock, or mandate a hard reset to preclinical validation phases for major algorithmic changes. By adhering to this operational matrix, healthcare institutions can treat post-market deployment as a continuous, governed validation loop, ensuring that the AI system evolves safely and responsively without compromising clinical integrity[141,149].

Strengths and limitations. The strength of the phase lies in its ability to accommodate the evolving nature of medical AI systems, shifting evaluation from static validation to dynamic monitoring[141]. Through cross-phase integration, cross-version tracking, and cross-context adaptation, it can detect risks and performance changes arising from algorithm updates, data shifts, or deployment-environment variation, thereby ensuring sustained effectiveness, safety, and fairness in real-world settings[149]. The framework also offers regulators an operational pathway for lifecycle management, enhancing transparency and control over AI approval, updating, and post-market surveillance[27]. However, its implementation depends on high-quality data infrastructure, continuous monitoring resources, and cross-departmental coordination, which may increase organizational and operational

Table 2 | A lifecycle-management matrix for dynamic AI evaluation: from monitoring to adaptation

MLOps lifecycle loop	Clinical & operational objectives	Illustrative metrics (Task-dependent)	Action and transition rules
Step A. Pre-deployment governance (PCCP Setup)	Establish operational boundaries, lock reference standards, and initialize traceability before real-world deployment.	Model registry (unique version IDs). Prespecified non-inferiority margins. Defined trigger thresholds for SSR.	Transition to live deployment: allowed only if static baseline validation meets prespecified gating criteria.
Step B. Active clinical surveillance	Monitor continuous real-world data streams to detect deviations from the baseline validation cohort.	Data drift: shifts in patient case-mix and heterogeneity in data-generation or acquisition processes. Behavioral drift: changes in clinician override rates or diagnostic confidence. Calibration drift: Increase in Brier score or ECE.	Continuous loop: if metrics remain within safety margins, continue routine real-world monitoring or silent background testing without intervention.
Step C. Trigger evaluation (Safety and drift alerts)	Detect critical performance decay or unsafe human–AI cognitive coupling requiring intervention.	Calibration drift: increase in Brier score or ECE. Safety alerts: e.g., >10% drop in sensitivity for specific high-risk subgroups.	Divergent actions: Minor drift: trigger local recalibration. Critical safety breach: trigger immediate rollback to the previous version.
Step D. Adaptation and recalibration	Safely execute adaptive updates (e.g., threshold tuning, cohort-specific retraining) to restore performance	Statistical power and sample size required for the new local cohort. Delta-performance (ΔAUC, Δspecificity) compared to the prior version.	Loop Re-entry vs. Hard Reset: Minor threshold tuning: deploy locally via "micro-trial" validation. Major algorithmic change: mandate return to static preclinical validation for full re-evaluation.
Step E. Post-update governance	Ensure transparent version transitions, mitigate "transition shock," and maintain accountability.	Immutable audit logs of parameter edits. User-notification records detailing the scope of the update. Updated device labeling.	Loop closure: lock the newly updated model version and return to Step B (active clinical surveillance) to begin a new lifecycle loop.

MLOps machine learning operations, *PCCP* predetermined change control plans, *SSR* sample size re-estimation, *ECE* expected calibration error, *AI* artificial intelligence, *AUC* area under the curve.

burden[151]. The evaluation process must balance sensitivity with feasibility to avoid both over- and under-monitoring, and further work is needed to establish its scalability and cross-institutional applicability across diverse healthcare systems.

Phase-gating criteria and fall-back triggers

To enhance the operational clarity of the five-phase evaluation framework, we added a dynamic phase-gating matrix (Table 3) that integrates core evidence domains, illustrative gating considerations, and fall-back/re-evaluation triggers across all phases. These illustrative gating considerations serve as qualitative success criteria that define the conditions under which a model may advance to the next phase. Importantly, these criteria are intended as risk-stratified operational guidance rather than universal thresholds; final standards must be tailored to the model's intended clinical role, risk profile, and application context.

The framework also distinguishes between local and systemic fall-back triggers. Local failures, such as transient throughput collapse, isolated subgroup instability, or interface-induced variability, require re-evaluation within the adjacent phase. In contrast, systemic failures (e.g., safety-signal emergence, severe drift, or fairness degradation) necessitate returning to earlier phases for renewed data auditing, recalibration, uncertainty re-estimation, or repeat MRMC or interaction-level evaluation. This dynamic re-entry mechanism ensures that continuously evolving AI systems remain safe, reliable, and clinically effective throughout their lifecycle.

Importantly, although Table 3 provides illustrative gating considerations, these are intended to guide rather than prescribe phase progression. Because this framework is designed to accommodate diverse AI applications, it deliberately does not specify universal numerical thresholds. Instead, phase-gating criteria should be prospectively defined according to the intended clinical role of the AI system, its associated patient safety risk, and the objectives of each evaluation phase[152]. In practice, progression decisions should integrate multiple complementary domains, including: (1) the intended clinical role (e.g., screening, triage, diagnosis, prognosis, or workflow support); (2) the potential consequences of incorrect predictions and the corresponding clinical risk level; (3) predefined minimum acceptable performance criteria established during study design, including prespecified superiority or non-inferiority hypotheses where appropriate[18,113]; (4) robustness across clinically relevant patient subgroups, healthcare sites, and acquisition conditions; (5) engineering performance requirements, such as latency, reliability, system availability, and interoperability[109]; and (6) applicable regulatory guidance, specialty-specific recommendations, or expert consensus. Accordingly, the gating considerations presented in Table 3 should be interpreted as illustrative operational guidance rather than mandatory thresholds, and quantitative decision criteria should be prespecified whenever possible and justified according to the specific clinical context.

We specifically highlight two critical dimensions that are frequently overlooked in conventional evaluations: uncertainty quantification and calibration. Integrated as core evaluation axes throughout the entire framework, they directly address the stringent safety and risk-stratification requirements of clinical decision-support systems.

Calibration is conceptualized in this framework as a task-dependent core evaluation dimension. Its primary function is to assess the concordance between model-output probabilities and actual event frequencies, making it indispensable for probabilistic prediction models, risk-stratification systems, and triage tools. In Phase 1, calibration evaluates the reliability and cross-subgroup stability of output probabilities, quantified by standard metrics such as reliability plots, Brier scores, and ECE[64,153]. In Phase 2, the focus shifts to calibration stability and drift within real-world clinical data streams. By Phase 3, calibration extends beyond a mere statistical property of the model; it directly dictates the recalibration of clinician trust, human–AI confidence alignment, and the ability to mitigate automation bias. Entering Phase 4, calibration evaluation formally transitions toward clinical utility. Here, DCA is explicitly incorporated to quantitatively assess whether a calibrated model provides genuine net clinical benefit across varying decision thresholds[41]. In Phase 5, calibration drift serves as a primary early-warning signal for active post-market surveillance and triggers PCCPs[145]. Notably, calibration is not a mandatory primary endpoint for pure detection, segmentation, or retrieval tasks, as these systems can still demonstrate clinical value by improving diagnostic efficiency or optimizing workflows despite limited probability calibration.

Table 3 | Illustrative phase-gating criteria and fall-back triggers across the five-phase evaluation framework

Transition and methods	Core evidence domains	Illustrative gating considerations	Fall-back/Re-evaluation triggers
Phase 1 → Phase 2 (Technical to Operational)	Diagnostic performance (AUC, accuracy, sensitivity, specificity); calibration and uncertainty; clinical utility (e.g., DCA); Subgroup robustness and fairness; site heterogeneity	Diagnostic performance within a prespecified acceptable range for the clinic, no subgroup-specific or site-specific collapse. Calibration and uncertainty quantification acceptable for predictive tasks. DCA shows positive net clinical benefit across relevant risk thresholds. AI-by-site interaction within acceptable limits.	Performance outside acceptable range; significant subgroup degradation; calibration instability; severe data drift; poorly validated or unreliable uncertainty estimation. Action: Return to model refinement/ retraining.
Phase 2 → Phase 3 (Operational to Interaction)	Baseline engineering robustness (uptime, single-user latency, nominal throughput, failure rate); scale-up engineering robustness (network latency under load, concurrent batch throughput, compute-resource ceilings); uncertainty; subgroup robustness; diagnostic performance; Calibration stability	Engineering performance meets prespecified thresholds under both nominal and peak-load conditions (no scale-up degradation, no compute-ceiling saturation). Model behavior remains stable under real data streams with acceptable subgroup drift; Pipeline stability is confirmed; and compute cost remains within operational limits.	System instability, unacceptable latency, or incompatibility with real-time data streams; performance degradation under concurrent load (throughput collapse, queue buildup, or resource saturation); or compute-ceiling violations or non-deterministic output behavior. Action: Return to Phase 2 engineering optimization.
Phase 3 → Phase 4 (Interaction to clinical)			
Standalone benchmarking	Diagnostic performance; efficiency; subgroup robustness; site heterogeneity	The AI's standalone diagnostic performance should be at least comparable to the clinician tier corresponding to its intended clinical role.	Performance substantially below the intended tier.
Exploratory paired design	Diagnostic performance; efficiency; confidence; subgroup robustness; site heterogeneity	AI-assisted readings show directionally consistent cognitive benefit (improved accuracy/confidence, reduced reading time or error rates) without harmful automation bias.	Inconsistent or harmful effects; unacceptable reader variability; interface-induced bias. Action: Redesign user interface/experience.
MRMC	Diagnostic performance; reader and case variability; subgroup robustness; site heterogeneity	AI-assisted mode demonstrates statistically and clinically meaningful improvement across readers and cases, with acceptable variability.	No net benefit; inconsistent effects across reader groups; excessive variability. Action: Return to Phase 3 for optimization.
Phase 4 → Phase 5 (Clinical to Real-world)			
Diagnostic RCT	Diagnostic performance; efficiency; diagnostic pathways; clinical utility; uncertainty; subgroup robustness; site heterogeneity	Clinically meaningful improvement in diagnostic performance or efficiency without increasing false-negative or false-positive harms; stable across sites/subgroups; DCA confirms clinical utility over standard care.	No diagnostic benefit, increases diagnostic errors, or produces inconsistent effects across sites or subgroups; negative net benefit in clinical utility assessment. Action: Return to Phase 3 for interface or workflow redesign.
Clinical-effectiveness RCT	Patient-level outcomes; safety and adverse events; health economics; clinical utility; site/ subgroup consistency	Improvement in patient outcomes (e.g., morbidity, time-to-treatment) with acceptable safety.	Primary endpoint not met; safety concerns; heterogeneous treatment effects. Action: refine model/workflow and repeat evaluation.
SW-CRT	System-level impact; workflow efficiency; implementation fidelity; safety signals; site heterogeneity	Consistent system-level improvement (reduced delays, improved throughput, fewer missed cases) across clusters with stable fidelity.	Cluster-specific negative effects, implementation failure, or safety drift during rollout. Action: pause rollout, conduct root-cause analysis, and revert affected clusters.
Phase 5 Real-world integration			
Single-arm prospective pre-post study	Pre-post change in workflow (diagnostic efficiency, diagnostic pathways, clinician behavior); early safety signals; subgroup performance checks	Post-implementation improvement in predefined metrics without new safety concerns.	Worsening of key metrics, unexpected error-mode shifts, or emerging safety concerns. Action: Pause deployment and return to Phase 3/4.
Real-world effectiveness study	Effectiveness in broad populations; heterogeneity of treatment effect; real-world safety; site heterogeneity; subgroup-specific effectiveness; fairness drift	Stable effectiveness across diverse populations and sites; acceptable safety; no harmful subgroup effects.	Site-specific performance collapse, harmful subgroup effects, or safety events. Action: Restrict deployment scope or return to controlled evaluation.
Implementation research	Implementation fidelity; Human–AI interaction patterns; Acceptability & usability; Workflow integration; Monitoring of workflow-level safety signals	High fidelity; stable human–AI interaction; acceptable adoption; no workflow disruption.	Low implementation fidelity, workflow breakdown, or negative human–AI interaction patterns. Action: Redesign workflow/interface.
Post-market active surveillance	Performance drift; calibration drift; uncertainty drift; safety-signal detection; Site-specific anomalies; control-chart monitoring outputs (CUSUM/EWMA/Shewhart)	No significant drift; stable performance and fairness across sites and time. All indicators within predefined alert thresholds (e.g., ≤5% AUROC drop, ≤10% sensitivity drop)*.	Drift beyond limits; safety events; anomalies. Action: Trigger PCCP-aligned update or restrict deployment. Control-chart rule violations (e.g., 2-of-3 beyond 2σ*, run-rule violations). Trigger version rollback/temporary suspension/PCCP-aligned update.

Table 3 (continued) | Illustrative phase-gating criteria and fall-back triggers across the five-phase evaluation framework

Transition and methods	Core evidence domains	Illustrative gating considerations	Fall-back/Re-evaluation triggers
Dynamic evaluation architecture	Interim analyses; sample size re-estimation; adaptive platform trials; versioned evaluation; drift detection; safety monitoring	Adaptive modifications proceed only when prespecified information thresholds are met, decision rules satisfied, and statistical validity preserved in alignment with intended use and PCCP when applicable.	Unstable interim analyses; violation of decision rules; emergent safety events after version update. Action: Suspend adaptive modifications and revert to fixed-design or earlier phase.

*Note: Illustrative examples only. These values are not recommended universal thresholds and should be defined according to the intended clinical application, deployment environment, monitoring strategy, and acceptable patient safety risk.

AI artificial intelligence, *AUC* area under curve, *DCA* decision-curve analysis, *MRMC* multi-reader multi-case, *RCT* randomized controlled trial, *SW-CRT* stepped-wedge cluster randomized controlled trial, *RWS* real-world effectiveness study, *IR* implementation research, *CUSUM* cumulative sum chart, *EWMA* exponentially weighted moving average, *PCCP* predetermined change control plans.

Uncertainty characterization may serve as a potential safety mechanism; however, its ability to indicate prediction unreliability or detect out-of-distribution cases is not guaranteed and must itself be empirically validated in the intended clinical context. In Phase 1, uncertainty evaluation should assess whether uncertainty estimates appropriately identify prediction unreliability, including heterogeneous data, edge cases, and out-of-distribution scenarios, and should explicitly validate the performance of uncertainty estimation rather than assuming its reliability. Phase 2 further evaluates the stability and reliability of uncertainty estimates under noisy real-world data streams, including their ability to detect distribution shifts without generating excessive false alarms[154,155]. In Phase 3, uncertainty display (e.g., confidence intervals or warning prompts) should be evaluated as a human-factor intervention rather than assumed to improve decision-making. Evaluation should examine how clinicians interpret and respond to uncertainty information, including its effects on confidence calibration, reliance behavior, automation bias, and diagnostic performance[156,157]. In Phase 4, appropriately validated uncertainty estimates may support escalation strategies and conservative decision-making for high-risk cases[155,156]. Finally, in Phase 5, anomalous fluctuations in predictive uncertainty may provide early signals of distribution shifts or fairness degradation and should be monitored alongside overall model performance to inform lifecycle re-evaluation and dynamic recalibration[145,154].

Discussion

Recent guidelines, including FUTURE-AI, SALIENT, CANAIRI, and the FDA GMLP, have laid a vital foundation for medical AI, though they primarily operate at conceptual, organizational, or regulatory levels. For example, FUTURE-AI offers 30 best practices built on six core principles for trustworthy AI[32]; SALIENT maps out end-to-end implementation tasks linked to reporting standards[60]; CANAIRI expands traditional silent trials into broader sociotechnical translational trials[48]; and the FDA GMLP outlines ten high-level principles for machine learning development[109]. However, these works generally stop short of providing prescriptive guidance on selecting specific study designs, estimating sample sizes, and methodologically handling continuous algorithmic iteration. As detailed in Supplementary Table 3, our proposed framework bridges the gap between these foundational principles and actual clinical execution by serving as a methodological blueprint. It advances the field by systematically mapping 18 actionable study designs across a five-phase clinical evaluation gradient. Furthermore, rather than treating evaluation as a static, pre-deployment hurdle, our approach incorporates a dynamic evaluation architecture. This allows researchers to methodologically accommodate the continuous iteration of AI models, directly aligning evidence generation with emerging regulatory mechanisms like PCCPs.

Beyond medical-AI-specific guidelines, several broader machine learning and maturity-model frameworks also inform the development and deployment of AI systems. CRISP-ML(Q) provides a structured process model for ML quality management[158], while OPTICA focuses on operational readiness and organizational workflows for AI deployment[19]. Maturity-level frameworks such as TRL, MDRL, and ML-TRL assess system readiness from engineering or regulatory perspectives[20,159], but they do not define validation pathways specific to adaptive, data-driven clinical AI systems. Moreover, these frameworks do not specify how medical AI should progress through clinically grounded, evidence-based validation phases, nor do they offer prescriptive guidance on study-design selection, sample-size estimation, multi-center heterogeneity, or dynamic model iteration. Importantly, none of these frameworks articulate concrete fall-back pathways or re-evaluation mechanisms when predefined criteria are not met. In contrast, our framework introduces an explicit phase-gating structure with clearly defined success criteria and both local and systemic fall-back triggers (Table 3). This structured re-entry mechanism, absent from existing frameworks, provides an auditable pathway for returning to earlier phases when performance degradation, safety signals, or fairness drift emerge, thereby supporting lifecycle-long safety assurance for continuously evolving medical AI systems. A detailed comparison is provided in Supplementary Table 4.

Clinical evaluation of medical AI is undergoing a profound paradigm shift, from a narrow "performance-metric competition" to a system-level model of lifecycle governance. Building on the five-phase evaluation framework and its overarching dynamic evaluation architecture developed in this study, we argue that future evaluations should prioritize six core domains.

Evaluating medical AI requires distinguishing between a constant diagnostic ground truth and evolving phase-specific evaluation endpoints (Supplementary Table 5)[16,112,113]. To validly track core algorithm performance (e.g., discrimination, calibration) across phases, the reference standard must consistently rely on biopsy-proven pathology or rigorously re-adjudicated senior-expert consensus, expressly avoiding unverified historical clinical labels[160]. This constant diagnostic ground truth underpins Phase 1, Phase 3, and the diagnostic performance evaluations within Phases 2 and 4. However, the overarching evaluation targets expand across the lifecycle. While Phases 1 and 3 strictly assess diagnostic accuracy against this constant truth, Phase 2 introduces additional operational metrics by comparing AI outputs to real-time clinical judgments. Furthermore, Phase 4 expands the evaluation focus to tangible patient outcomes in clinical-effectiveness RCTs[161], and Phase 5 shifts to evaluating post-market performance and system-level effectiveness. Consequently, valid cross-phase comparisons demand strict construct alignment. For instance, tracking calibration drift from Phase 1 to Phase 2 is valid only when calculated against the identical pathology/expert ground truth. Conversely, it is methodologically invalid to equate high operational agreement with clinicians in Phase 2 with actual patient benefit in Phase 4[22].

The focus must shift from validating performance ceilings to identifying failure boundaries. Medical AI introduces complex forms of bias, such as spectrum bias, data drift, and automation bias[22,162], that extend beyond the scope of traditional diagnostic-test evaluation. These biases evolve across validation stages, manifesting as selective data bias in early phases and transforming into systemic bias after real-world deployment[163]. The key bias types across the five-phase evaluation framework and its dynamic evaluation architecture are summarized in Supplementary Table 6. Algorithm vigilance is therefore essential. Stress testing and pitfall-case evaluation can proactively reveal performance

degradation under extreme or atypical conditions, enabling a pharmacovigilance-like monitoring loop that safeguards safety across the AI lifecycle.

Evaluation must move from measuring AI performance to assessing the performance of the human–AI system. AI now functions as a multi-role partner within clinical workflows, a parallel assistant, second reader, triage tool, or quasi-autonomous decision system[67,115,164]. Traditional metrics such as AUC cannot capture dimensions such as workflow efficiency, missed-case recovery, or intervention delay. A core challenge is cognitive coupling: clinicians' decision behavior, trust calibration, and vigilance may be reshaped by AI over time[67,165,166]. Future evaluation should dissect these interaction dynamics, address risks of long-term de-skilling, and incorporate clinician satisfaction and cognitive load as core outcomes to support a transparent and adaptive human–AI ecosystem.

Evaluation must transition from snapshot validation to traceable, dynamically updated evidence. The fixed-intervention logic is structurally misaligned with the continuous iteration of AI systems[141]. Dynamic evaluation architecture is therefore critical to bridging the gap between evidence-generation cycles and algorithm-evolution speed. Regulatory science is also shifting: frameworks such as the PCCPs signal a move from fixed-version confirmation to closed-loop dynamic governance[145]. Meanwhile, to address the "black-box" nature of AI, evaluation must incorporate quantitative assessment of explainability to ensure alignment with clinical reasoning, and enforce fairness auditing as a baseline requirement for safeguarding equity[167]. Consistent with this dynamic perspective, evidence generation itself is not confined to a linear sequence of phases. Implementation research and sociotechnical context scoping may begin well before Phase 4 randomized trials, while real-world system-level insights from Phase 5 can iteratively refine earlier-phase simulation parameters, workflow assumptions, and trial designs[168]. This legitimate parallelism and bidirectional feedback ensure that evidence generation remains adaptive, context-aware, and aligned with real-world deployment conditions.

The proposed five-phase evaluation framework operationalizes key dimensions of ethical AI governance and provides a practical pathway for meeting global regulatory expectations. Its phase-gating criteria and fall-back triggers directly implement core principles emphasized in international standards such as the OECD AI Principles[24] and the EU AI Act[25], including robustness, fairness, transparency, human oversight, and lifecycle risk management. Calibration stability, uncertainty characterization, subgroup robustness, and safety-signal detection are embedded as mandatory evidence domains across phases, ensuring that ethical and safety considerations are evaluated continuously rather than treated as static pre-deployment requirements[109]. The dynamic re-entry mechanism further aligns with emerging legal requirements for continuous monitoring, post-market surveillance, and controlled model updates for adaptive AI systems[145]. By integrating these ethical and regulatory expectations into a structured, auditable evaluation gradient, the framework enables developers and regulators to generate the evidence needed to demonstrate compliance throughout the AI lifecycle.

Evaluation must extend beyond demonstrating clinical efficacy to establishing economic viability. Regulatory clearance and clinical effectiveness alone are no longer sufficient for large-scale adoption of medical AI; most HTA bodies now require evidence on cost-effectiveness, budget impact[169], and value-of-information[170] before approving widespread deployment or public reimbursement[171]. In our framework, these health-economic components are incorporated within the later phases (Phases 4 and 5) as essential elements of real-world implementation and real-world integration. However, full economic modeling depends on highly specific AI applications, local cost structures, and regional clinical pathways[172]. Accordingly, the present manuscript focuses on establishing the overarching conceptual and methodological architecture rather than performing application-specific economic simulations. Future applications of this framework to concrete AI systems will need to integrate these HTA elements to bridge the gap between clinical evidence generation and institutional procurement[171].

Despite its contributions, this study has several limitations. First, as a conceptual framework, its generalizability and operational feasibility require empirical validation across diverse clinical domains, data ecosystems, and healthcare systems. Second, the phase-gating thresholds, fall-back triggers, and dynamic re-entry mechanisms are grounded in theoretical reasoning and emerging regulatory trends, but lack quantitative calibration across institutions and real-world settings. Third, the framework is primarily oriented toward diagnostic and prognostic SaMD-type systems; further adaptation is needed for generative models, continuously learning algorithms, and non-diagnostic AI applications such as workflow optimization tools. Fourth, the reproducibility and standardization of human–AI interaction assessments (Phase 3) across heterogeneous clinical environments remain to be established through both experimental and real-world studies. Finally, while the framework integrates major ethical and regulatory governance directives, it does not fully encompass the broader macro-economic, organizational, and fine-grained sociotechnical factors that shape practical AI deployment, which future research should integrate more comprehensively.

Clinical evaluation of medical AI is shifting from fragmented technical validation to a system-level model of lifecycle governance. Using a five-phase evaluation framework and dynamic evaluation architecture encompassing eight core methods and 18 study designs, this study maps the evidence-generation pathway from algorithmic cold start to deep clinical integration, emphasizing that evaluation is not a single entry-point test but a staged, co-evolving evidence system shaped by context, actors, and technology (Table 4). Beyond proposing a structured gradient framework, this study articulates the 'dilution mechanism' of algorithmic efficacy in real-world settings through a tripartite framework of "physical, consensus, and functional truth". Furthermore, by incorporating adaptive design, the framework provides a potential pathway to reconcile evidence generation with continuous model evolution. Finally, the value of medical AI lies not in constructing a flawless "digital oracle", but in building a safe, fair, trustworthy, and continuously evolving human–AI system. A full-lifecycle evidence framework is therefore essential to support clinical decision-making and enable the safe, effective, and sustainable integration of medical AI into complex real-world healthcare settings.

Methods

This study developed the five-phase evaluation framework and dynamic lifecycle architecture through an iterative conceptual-modeling process. The framework was refined by integrating insights from methodological literature, existing AI evaluation guidelines, regulatory documents, and established principles from diagnostic test research, clinical-trial methodology, software-engineering robustness testing, and human–AI interaction studies. Concepts identified from these sources were organized into coherent evaluation structure, which informed the definition of phase objectives, appropriate study designs, and evidence requirements. Illustrative phase-gating criteria and fall-back triggers were developed by synthesizing principles from lifecycle risk management and safety-critical system monitoring. The overall structure and terminology were iteratively reviewed to ensure methodological clarity, operational feasibility, and alignment with emerging regulatory expectations for adaptive medical AI systems.

Table 4 | Summary of characteristics of the five-phase evaluation framework for the clinical evaluation of medical artificial intelligence

Phase	Study design	Regulatory positioning	Environmental /randomization /intervention	Recommended reporting guidelines	Core objective
Phase 1. Technical validation	Multicenter retrospective external validation	Pre-market technical qualification	Laboratory (in silico); Retrospective; No randomization; No intervention (pure algorithmic testing)	STARD-AI (for diagnostic accuracy)/ TRIPOD + AI (for individual risk prediction)	Establishing data robustness across different centers, devices, and populations
Phase 2. Operational robustness	Shadow-mode /silent trial	Pre-market deployment readiness	Real-world; Prospective; No randomization; No intervention	N/A (Local MLOps logging protocols)	Evaluating engineering stability and data drift risks within real hospital data flows.
Phase 3. Controlled interaction	Standalone benchmarking; Paired design; MRMC	Pre-registration synergistic gain confirmation	Laboratory (human–AI interaction); Retrospective; Paired design / MRMC: Randomization with washout period; AI intervention	DECIDE-AI (primary)/STARD-AI (complementary for joint accuracy)	Defining AI's role and isolating the "net incremental gain" of AI on physician diagnostic behavior
Phase 4. Clinical evidence	Diagnostic RCT; Clinical effectiveness RCT; SW-CRT	Registration-enabling clinical evidence	Highly controlled, ideal clinical setting; Prospective; Randomization; AI intervention	SPIRIT-AI (for protocol design) /CONSORT-AI (for trial reporting)	Evaluating whether AI improves patient outcomes and clinical pathways
Phase 5. Real-world integration	Single-arm prospective pre-post study; RWS; IR; Post-market active surveillance	Post-market system-level performance governance	Real-world; Open environment; Weak or natural intervention; No randomization	STROBE/STARI/ Emerging PMS standards	Evaluating the long-term effectiveness, safety, and equity of AI in the real world

TRIPOD + AI transparent reporting of a multivariable prediction model for individual prognosis or diagnosis + artificial intelligence, *STARD-AI* Standards for reporting diagnostic accuracy—artificial intelligence, *DECIDE-AI* developmental and exploratory clinical investigations of decision-support systems driven by artificial intelligence; *SPIRIT-AI* standard protocol items: recommendations for interventional trials—artificial intelligence, *CONSORT-AI* Consolidated standards of reporting trials–artificial Intelligence, *STROBE* strengthening the reporting of observational studies in epidemiology, *STARI* standards for reporting implementation studies, *PMS* post-market surveillance.
MRMC multi-reader multi-case, *RCT* randomized controlled trial, *SW-CRT* stepped-wedge cluster randomized controlled trial, *RWS* real-world effectiveness study, *IR* implementation research, *AI* artificial intelligence, *PCCP* predetermined change control plan, *MLOps* machine learning operations.

Data availability

No datasets were generated or analysed during the current study.

Code availability

No code was generated or used in this study.

Received: 24 April 2026; Accepted: 11 August 2026;
Published online: 03 September 2026

References

1. Bolton, W. J. et al. The impact of artificial intelligence-driven decision support on uncertain antimicrobial prescribing: a randomised, multimethod study. *Lancet Digit. Health* **7**, 100912 (2025).
2. Yesankar, P., Puri, C. & Gote, P. M. AI-powered clinical decision support systems (CDSS): challenges, benefits, applications, and future directions. In *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)* 1192–1197 (IEEE, 2025).
3. Alowais, S. A. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
4. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
5. Tschandl, P. et al. Human-computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
6. Lin, L. et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology* **291**, 677–686 (2019).
7. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
8. Marra, A. et al. Artificial intelligence entering the pathology arena in oncology: current applications and future perspectives. *Ann. Oncol.* **36**, 712–725 (2025).
9. Shaktah, L. A. et al. Application of artificial intelligence and digital tools in cancer pathology. *Lancet Digit. Health* **7**, 100933 (2025).
10. Steiner, D. F., Chen, P. C. & Mermel, C. H. Closing the translation gap: AI applications in digital pathology. *Biochim. Biophys. Acta Rev. Cancer* **1875**, 188452 (2021).
11. Overgaard, S. M. et al. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. *NPJ Digit. Med.* **6**, 218 (2023).
12. Reyna, M. A., Nsoesie, E. O. & Clifford, G. D. Rethinking algorithm performance metrics for artificial intelligence in diagnostic medicine. *JAMA* **328**, 329–330 (2022).
13. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
14. Chouffani El Fassi, S. et al. Not all AI health tools with regulatory authorization are clinically validated. *Nat. Med.* **30**, 2718–2720 (2024).
15. Talianu, A. et al. Artificial intelligence for improving decision-making in bacterial infection management: a narrative review. *J. Antimicrob. Chemother.* **81**, dkaf470 (2026).
16. Vasey, B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).
17. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
18. Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **31**, 3283–3289 (2025).
19. Dagan, N. et al. Evaluation of AI solutions in health care organizations — the OPTICA tool. *NEJM AI* **1**, AIcs2300269 (2024).

20. Lavin, A. et al. Technology readiness levels for machine learning systems. *Nat. Commun.* **13**, 6039 (2022).
21. Müller, K. -R. Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *Mach. Lear. Know. Extr.* **3**, 392–413 (2021).
22. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. **17**, 195 (2019).
23. Shah, N. H., Milstein, A. & Bagley Ph, D. S. Making machine learning models clinically useful. *JAMA* **322**, 1351–1352 (2019).
24. Organisation for Economic Co-operation and Development. Recommendation of the council on artificial intelligence. 2026 (2024).
25. European Parliament and Council of the European Union. Regulation (EU) of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence. 2026 (2024).
26. Yuan, H. Toward real-world deployment of machine learning for health care: External validation, continual monitoring, and randomized clinical trials. *Health Care Sci.* **3**, 360–364 (2024).
27. U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. FDA Guidance Document. 2026 (2025).
28. Xue, P. et al. Deep learning enabled liquid-based cytology model for cervical precancer and cancer detection. *Nat. Commun.* **16**, 3506 (2025).
29. Hu, B. et al. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. *Lancet Digit. Health* **6**, e261–e271 (2024).
30. Clift, A. K. et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. *BMJ* **381**, e073800 (2023).
31. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
32. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).
33. Yu, A. C., Mohajer, B. & Eng, J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radio. Artif. Intell.* **4**, e210064 (2022).
34. Bertens, L. C. et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med.* **10**, e1001531 (2013).
35. Knottnerus, J. A. & Muris, J. W. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J. Clin. Epidemiol.* **56**, 1118–1128 (2003).
36. Fanti, S. & Lalumera, E. Of standard of reference and accuracy: the problem of truth in imaging. *Eur. J. Nucl. Med. Mol. Imaging* **43**, 52–54 (2016).
37. O'Sullivan, J. W., Banerjee, A., Heneghan, C. & Pluddemann, A. Verification bias. *BMJ Evid. Based Med.* **23**, 54–55 (2018).
38. Harden, M. & Friede, T. Sample size calculation in multi-centre clinical trials. *BMC Med. Res. Methodol.* **18**, 156 (2018).
39. Zhou, H. H. et al. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proc. Natl. Acad. Sci. USA* **115**, 1481–1486 (2018).
40. Arun, S. et al. Systematic scoping review of external validation studies of AI pathology models for lung cancer diagnosis. *NPJ Precis. Oncol.* **9**, 166 (2025).
41. Efthimiou, O. et al. Developing clinical prediction models: a step-by-step guide. *BMJ* **386**, e078276 (2024).
42. Chen, F., Lu, H. L. & Simonetti, A. Practical statistical considerations for the clinical validation of AI/ML-enabled medical diagnostic devices. Preprint at arXiv:2303.05399v1 (2023).
43. Topff, L. et al. Is the generalizability of a developed artificial intelligence algorithm for COVID-19 on chest CT sufficient for clinical use? Results from the International Consortium for COVID-19 Imaging AI (ICOVAI). *Eur. Radiol.* **33**, 4249–4258 (2023).
44. Cabitza, F. et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput. Methods Prog. Biomed.* **208**, 106288 (2021).
45. Marconi, L. & Cabitza, F. Show and tell: a critical review on robustness and uncertainty for a more responsible medical AI. *Int. J. Med. Inform.* **202**, 105970 (2025).
46. Tonekaboni, S. et al. How to validate machine learning models prior to deployment: silent trial protocol for evaluation of real-time models at ICU. In *ACM Conference on Health, Inference, and Learning* (2022).
47. Kwong, J. C. C. et al. The silent trial - the bridge between bench-to-bedside clinical AI applications. *Front. Digit. Health* **4**, 929508 (2022).
48. McCradden, M. D. et al. CANAIRI: the collaboration for translational artificial intelligence trials in healthcare. *Nat. Med.* **31**, 9–11 (2025).
49. McCradden, M. D., Stephenson, E. A. & Anderson, J. A. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat. Med.* **26**, 1325–1326 (2020).
50. Flack, V., Afifi, A., Lachenbruch, P. & Schouten, H. Sample size determinations for the two rater kappa statistic. *Psychometrika* **53**, 321–325 (1988).
51. Kraemer, H. C., Bloch, D. A., Donner, A. & Eliasziw, M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat. Med.* **11**, 1511–1519 (1994).
52. Daniel, W. W. & Cross, C. L. *Biostatistics: A Foundation for Analysis in the Health Sciences* (John Wiley & Sons, 2018).
53. Cochran, W. G. *Sampling Techniques* (John Wiley & Sons, 1977).
54. Lwanga, S. K. & Lemeshow, S. *Sample Size Determination in Health Studies: A Practical Manual* (World Health Organization, 1991).
55. Hanley, J. A. & Lippman-Hand, A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* **249**, 1743–1745 (1983).
56. Jovanovic, B. D. & Levy, P. S. A look at the rule of three. *Am. Stat.* **51**, 137–139 (1997).
57. Newman, T. B. If almost nothing goes wrong, is almost everything all right? Interpreting small numerators. *JAMA* **274**, 1013 (1995).
58. Casella, G. & Berger, R. L. *Statistical Inference* (Chapman and Hall/CRC, 2024).
59. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413 (1934).
60. van der Vegt, A. H. et al. Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework. *J. Am. Med. Inf. Assoc.* **30**, 1503–1515 (2023).
61. Obadinma, S. et al. The FAIIR conversational AI agent assistant for youth mental health service provision. *NPJ Digit. Med.* **8**, 243 (2025).
62. Corbin, C. K. et al. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *J. Am. Med. Inf. Assoc.* **30**, 1532–1542 (2023).
63. Jiang, B. et al. Validation of a deep learning model for traumatic brain injury detection and NIRIS grading on non-contrast CT: a multi-reader study with promising results and opportunities for improvement. *Neuroradiology* **65**, 1605–1617 (2023).
64. Riley, R. D. et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* **384**, e074820 (2024).
65. Ye, Z. et al. Deep learning algorithms for melanoma detection using dermoscopic images: a systematic review and meta-analysis. *Artif. Intell. Med.* **155**, 102934 (2024).

66. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
67. Ye, Z. et al. Deep learning-assisted system improves practical effects in cervical cytopathology diagnosis: a comparative study of reading modes. *Mod. Pathol.* **38**, 100882 (2025).
68. Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
69. Ye, Z. C., Xue, P., Qiao, Y. L. & Jiang, Y. Challenges in application of artificial intelligence in healthcare field and response strategies. *Chin. J. Epidemiol.* **45**, 1030–1038 (2024).
70. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
71. Obuchowski, N. A. Computing sample size for receiver operating characteristic studies. *Invest. Radio.* **29**, 238–243 (1994).
72. Chow, S., Wang, H. & Shao, J. *Sample Size Calculations in Clinical Research*. 2nd edn (Chapman & Hall/CRC, 2007).
73. Fleiss, J. L., Levin, B. & Paik, M .C. *Statistical Methods for Rates and Proportions*. 3rd edn (Wiley, 2003).
74. Xue, P. et al. Improving the accuracy and efficiency of abnormal cervical squamous cell detection with cytologist-in-the-loop artificial intelligence. *Mod. Pathol.* **36**, 100186 (2023).
75. Naeem, A. et al. Language artificial intelligence models as pioneers in diagnostic medicine? A retrospective analysis on real-time patients. *J. Clin. Med.* **14**, 1131 (2025).
76. Park, S. H. et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* **306**, 20–31 (2023).
77. Yang, B. et al. Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme. *J. Clin. Epidemiol.* **138**, 128–138 (2021).
78. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
79. Bussone, A., Stumpf, S. & Sullivan, D.O. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics* 160–169 (ACM, 2015).
80. Steiner, D. F. et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
81. Obuchowski, N. A. Sample size calculations in studies of test accuracy. *Stat. Methods Med. Res.* **7**, 371–392 (1998).
82. Shieh, G. Assessing individual equivalence in parallel group and crossover designs: exact test and sample size procedures. *PLoS ONE* **17**, e0269128 (2022).
83. Lin, M. et al. Deep learning system improved detection efficacy of fetal intracranial malformations in a randomized controlled trial. *NPJ Digit. Med.* **6**, 191 (2023).
84. Dratsch, T. et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **307**, e222176 (2023).
85. Kim, S. H. et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. *Radio. Med.* **130**, 555–566 (2025).
86. Administration, U.S.F.a.D. *Clinical Performance Assessment: Considerations for Computer-Assisted Detection (CADe) Devices Applied to Radiology Images and Radiology Device Data - Guidance for Industry and Food and Drug Administration Staff* (ed. Services, D.o.H.a.H.) (FDA, 2022).
87. U.S. Food and Drug Administration. iMRMC: software to do multi-reader multi-case statistical analysis of reader studies (RST24MD06.02). 2026 (2025).
88. Robert, D., Sathyamurthy, S. & Putha, P. MRMCsamplesize: an R package for estimating sample sizes for multi-reader multi-case studies. Preprint at *medRxiv* https://doi.org/10.1101/2023.09.25.23296069 (2023).
89. Obuchowski, N. A. & Bullen, J. Multireader diagnostic accuracy imaging studies: fundamentals of design and analysis. *Radiology* **303**, 26–34 (2022).
90. Chakraborty, D.P. *Observer Performance Methods for Diagnostic Imaging : Foundations, Modeling, and Applications with R-based Examples* (CRC Press, 2017).
91. Hillis, S. L. & Schartz, K. M. Multireader sample size program for diagnostic studies: demonstration and methodology. *J. Med. Imaging* **5**, 045503 (2018).
92. Capili, B. & Anastasi, J. K. An introduction to the crossover trial design. *Am. J. Nurs.* **124**, 40–43 (2024).
93. Obuchowski, N. A. Jr & Rockette, H. E. Jr Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an ANOVA approach with dependent observations. *Commun. Stat. Simul. Comput.* **24**, 285–308 (1995).
94. Hillis, S. L. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Stat. Med.* **26**, 596–619 (2007).
95. Hillis, S. L., Obuchowski, N. A., Schartz, K. M. & Berbaum, K. S. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Stat. Med.* **24**, 1579–1607 (2005).
96. Smith, B. J. & Hillis, S. L. Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proc. SPIE Int. Soc. Opt. Eng.* **11316**, 113160K (2020).
97. Hillis, S. L. & Schartz, K. M. Multireader Sample Size Program for Diagnostic Studies User Manual. *University of Iowa* (2018).
98. Hillis, S. L., Schartz, K. M. & Berbaum, K. S. OR-DBM MRMC 2.5 User Guide. *The University of Iowa* (2020).
99. Robert, D. MRMCsamplesize: sample size estimations for planning multi-reader multi-case (MRMC) studies without pilot data. *CRAN: Contributed Packages* (2023).
100. U.S. Food and Drug Administration. Statistical guidance on reporting results from studies evaluating diagnostic tests - guidance for industry and FDA staff. 2026 (2007).
101. Smith, B.J. Multi-reader multi-case analysis of variance for R getting started. 2026 (2023).
102. Guermazi, A. et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology* **302**, 627–636 (2022).
103. Obuchowski, N. A. & Rockette, H. E. Hypothesis-testing of diagnostic-accuracy for multiple readers and multiple tests-an ANOVA approach with dependent observations. *Commun. Stat. Simul. Comput.* **24**, 285–308 (1995).
104. Rockette, H. E. et al. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Acad. Radio.* **6**, 723–729 (1999).
105. Wu, Y. et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat. Med.* **31**, 3404–3413 (2025).
106. ClinicalTrials.gov. Point-of-care AI assistance and critical care outcomes: a randomized trial (POC-AI-ICU). 2026 (2025).
107. Sauro, K. M. et al. Enhanced recovery after surgery guidelines and hospital length of stay, readmission, complications, and mortality: a meta-analysis of randomized clinical trials. *JAMA Netw. Open* **7**, e2417310 (2024).
108. Plana, D. et al. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw. Open* **5**, e2233946 (2022).
109. US Food and Drug Administration. Good machine learning practice for medical device development: guiding principles. 2026 (2021).

110. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* **370**, m3164 (2020).
111. Guan, Z., Zeng, D., Li, H., Wong, T.Y. & Sheng, B. Can generative artificial intelligence empower target trial emulations? *Lancet Digit. Health* **8**, 100950 (2026).
112. Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K. & Calvert, M. J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* **370**, m3210 (2020).
113. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
114. Dijkgraaf, M. G. W., Haukoos, J. & Itani, K. M. F. Practical guide to design choice of randomized clinical trials in surgery. *JAMA Surg.* **157**, 1154–1155 (2022).
115. Fisches, Z. V. et al. Strategies for integrating artificial intelligence into mammography screening programmes: a retrospective simulation analysis. *Lancet Digit. Health* **6**, e803–e814 (2024).
116. Sanders, G. D. et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *JAMA* **316**, 1093–1103 (2016).
117. El Arab, R. A. & Al Moosa, O. A. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. *NPJ Digit. Med.* **8**, 548 (2025).
118. Farahani, S. & Liu, S. Reframing AI evaluation: from performance metrics to clinical impact. *Eur. J. Radiol. Artif. Intell.* **4**, 100051 (2025).
119. Debray, T. P. A. et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ* **380**, e071018 (2023).
120. Chu, R. et al. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Med. Res. Methodol.* **11**, 21 (2011).
121. Kahan, B. C. & Morris, T. P. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects?. *Stat. Med.* **32**, 1136–1149 (2013).
122. Campbell, M. K., Piaggio, G., Elbourne, D. R. & Altman, D. G. Consort 2010 statement: extension to cluster randomised trials. *BMJ* **345**, e5661 (2012).
123. Stefanos, R., Graziella, D. & Giovanni, T. Methodological aspects of superiority, equivalence, and non-inferiority trials. *Intern. Emerg. Med.* **15**, 1085–1091 (2020).
124. Wang, B., Wang, H., Tu, X. M. & Feng, C. Comparisons of superiority, non-inferiority, and equivalence trials. *Shanghai Arch. Psychiatry* **29**, 385–388 (2017).
125. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* **6**, e367–e373 (2024).
126. He, B. et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature* **616**, 520–524 (2023).
127. Zhou, Q. Noninferiority design in trials involving artificial intelligence interventions: a field needs to be improved. *Gastroenterology* **161**, 1072–1073 (2021).
128. Mauri, L. & D'Agostino, R. B. Sr Challenges in the design and interpretation of noninferiority trials. *N. Engl. J. Med.* **377**, 1357–1367 (2017).
129. U.S. Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness - guidance for industry. 2026 (2016).
130. El Arab, R.A. et al. Bridging the gap: from AI success in clinical trials to real-world healthcare implementation-a narrative review. *Healthcare* **13**, 701 (2025).
131. Kelshiker, M. A. et al. Triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (TRICORDER): design and rationale for a decentralised, real-world cluster-randomised controlled trial and implementation study. *BMJ Open* **15**, e098030 (2025).
132. Wilson, B. E. & Booth, C. M. Real-world data: bridging the gap between clinical trials and practice. *EClinicalMedicine* **78**, 102915 (2024).
133. Wieseler, B., Neyt, M., Kaiser, T., Hulstaert, F. & Windeler, J. Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy?. *BMJ* **380**, e073100 (2023).
134. Sherman, R. E. et al. Real-world evidence - what is it and what can it tell us?. *N. Engl. J. Med.* **375**, 2293–2297 (2016).
135. Sullivan, S. D. et al. Budget impact analysis—principles of good practice: report of the ISPOR 2012 budget impact analysis good practice II task force. *Value Health* **17**, 5–14 (2014).
136. Fenwick, E. et al. Value of information analysis for research decisions—an introduction: report 1 of the ISPOR value of information analysis emerging good practices task force. *Value Health* **23**, 139–150 (2020).
137. Kaack, L. H. et al. Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Change* **12**, 518–527 (2022).
138. Liu, F. & Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **22**, 287 (2022).
139. Lambert, J. et al. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. *Blood Adv.* **7**, 5680–5690 (2023).
140. Koski, E. et al. Towards responsible artificial intelligence in healthcare-getting real about real-world data and evidence. *J. Am. Med. Inf. Assoc.* **32**, 1746–1755 (2025).
141. Rosenthal, J. T., Beecy, A. & Sabuncu, M. R. Rethinking clinical trials for medical AI with dynamic deployments of adaptive systems. *NPJ Digit. Med.* **8**, 252 (2025).
142. Yu, C. et al. FRAME: feedback-refined agent methodology for enhancing medical research insights. In *Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2025).
143. Badani, A., de Moraes, F. Y., Vollmuth, P., Chung, C. & Mansouri, A. AI and innovation in clinical trials. *npj Digit. Med.* **8**, 683 (2025).
144. Pardoux, É & Kerasidou, A. Compliance with clinical guidelines and AI-based clinical decision support systems: implications for ethics and trust. *Sci. Eng. Ethics* **31**, 34 (2025).
145. 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. 2026 (2025).
146. U.S. Food and Drug Administration. Adaptive design clinical trials for drugs and biologics guidance for industry. 2026 (2019).
147. Webb, S. A. & Derde, L. P. G. Platform clinical trials for the efficient evaluation of multiple treatments. *JAMA.* **333**, 1166–1167 (2025).
148. Chiaruttini, M. V., Lorenzoni, G. & Gregori, D. Bayesian dynamic borrowing in group-sequential design for medical device studies. *BMC Med. Res. Methodol.* **25**, 78 (2025).
149. Feng, J. et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit. Med.* **5**, 66 (2022).
150. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
151. Maimaitiaili, M. et al. Artificial intelligence platform architecture for hospital systems: systematic review. *J. Med. Internet Res.* **27**, e79788 (2025).
152. International Medical Device Regulators Forum. software as a medical device: possible framework for risk categorization and corresponding considerations. 2026 (2014).

153. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
154. Ovadia, Y. et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proc. 33rd International Conference on Neural Information Processing Systems* Article 1254 (Curran Associates Inc., 2019).
155. Gawlikowski, J. et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).
156. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Med.* **4**, 4 (2021).
157. Yang, Q., Steinfeld, A., Rosé, C. & Zimmerman, J. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–13 (Association for Computing Machinery, 2020).
158. Studer, S. et al. Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *Mach. Learn. Knowl. Extr.* **3**, 392–413 (2021).
159. Seva, R. R., Tan, A. L. S., Tejero, L. M. S. & Salvacion, M. L. D. S. Multi-dimensional readiness assessment of medical devices. *Theor. Issues Ergon. Sci.* **24**, 189–205 (2022).
160. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
161. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
162. Chen, F., Wang, L., Hong, J., Jiang, J. & Zhou, L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J. Am. Med. Inf. Assoc.* **31**, 1172–1183 (2024).
163. Cross, J. L., Choma, M. A. & Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. *PLoS Digit. Health* **3**, e0000651 (2024).
164. Liu, F. et al. A foundational architecture for AI agents in healthcare. *Cell Rep. Med.* **6**, 102374 (2025).
165. Ye, Z., Lu, Q., Wang, J., Jiang, Y. & Xue, P. A quantitative study of pathologists' perceptions towards artificial intelligence-assisted diagnostic system. *PLoS Digit. Health* **4**, e0001052 (2025).
166. Monteith, S. et al. Artificial intelligence and deskilling in medicine. *Br. J. Psychiatry* 1–3 (2026).
167. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
168. Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M. & Stetler, C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med. Care* **50**, 217–226 (2012).
169. National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies. 2026 (2022).
170. Conrads-Frank, A. et al. Decision-analytic modeling for early health technology assessment of medical devices: a scoping review. *Ger. Med. Sci.* **20**, Doc11 (2022).
171. Kristensen, F. B. et al. Identifying the need for good practices in health technology assessment: summary of the ISPOR HTA Council Working Group report on good practices in HTA. *Value Health* **22**, 13–20 (2019).
172. Drummond, M.E., Sculpher, M.J., Torrance, G.W., O'Brien, B.J. & Stoddart, G.L. *Methods for the Economic Evaluation of Health Care Programmes* (Oxford Univ. Press, 2005).
173. Dwan, K., Li, T., Altman, D. G. & Elbourne, D. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ* **366**, l4378 (2019).
174. Liu, X. L., Wu, N. & Li, Y. F. Designs and appropriate choices for diagnostic test accuracy study. *Chin. J. Epidemiol.* **45**, 1705–1714 (2024).
175. Yu, D. et al. An introduction to the statistical analysis of stepped cluster randomized trials. *China J. Evid. Based Med.* **24**, 355–363 (2024).
176. Martinez-Gutierrez, J. C. et al. Automated large vessel occlusion detection software and thrombectomy treatment times: a cluster randomized clinical trial. *JAMA Neurol.* **80**, 1182–1190 (2023).
177. Hemming, K. & Taljaard, M. Extending a parallel cluster randomised trial into a stepped-wedge cluster randomised trial: implications for interpretation. *BMJ Glob. Health* **10**, e018087 (2025).
178. Eichner, F. A., Groenwold, R. H. H., Grobbee, D. E. & Oude Rengerink, K. Systematic review showed that stepped-wedge cluster randomized trials often did not reach their planned sample size. *J. Clin. Epidemiol.* **107**, 89–100 (2019).
179. Lee, K.A. et al. Real-world before-and-after evaluation of AI support for lung cancer diagnosis at three US lung nodule clinics. Preprint at *medRxiv* https://doi.org/10.1101/2025.11.20.25340677 (2025).
180. Wang, M. et al. Single-arm clinical trials: design, ethics, principles. *BMJ Support Palliat. Care* **15**, 46–54 (2024).
181. Eisemann, N. et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat. Med.* **31**, 917–924 (2025).
182. Smith, J. D. & Hasan, M. Quantitative approaches for the evaluation of implementation research studies. *Psychiatry Res.* **283**, 112521 (2020).
183. Taylor, L., Liang, H., Chen, C., Yang, X. & Xu, D. Evaluation frameworks in implementation science: a review of key features and practical applications. *Med. J. Peking. Union Med. Coll. Hospital* **16**, 1578–1586 (2025).
184. Theobold, M. EU postmarket surveillance: a deep dive into MDR compliance. 2026 (2025).
185. Ren, Y., Bertoldi, M., Fraser, A. G. & Caiani, E. G. Validation of CORE-MD PMS support tool: a novel strategy for aggregating information from notices of failures to support medical devices' post-market surveillance. *Ther. Innov. Regul. Sci.* **57**, 589–602 (2023).
186. Lin, R., Yang, Z., Yuan, Y. & Yin, G. Sample size re-estimation in adaptive enrichment design. *Contemp. Clin. Trials* **100**, 106216 (2021).
187. Jin, R., Wu, C. & Deng, Q. Optimal and efficient sample size re-estimation: a dynamic cost framework. Preprint at arXiv:2509.16636v2 (2025).
188. Adaptive platform trials: definition, design, conduct and reporting considerations. *Nat. Rev. Drug Discov.* **18**, 797–807 (2019).
189. Koenig, F. et al. Current state-of-the-art and gaps in platform trials: 10 things you should know, insights from EU-PEARL. *EClinicalMedicine* **67**, 102384 (2024).

Acknowledgements

This work was supported by the Beijing Natural Science Foundation of China (Grant No. 7264296), the National Natural Science Foundation of China (Grant No. 82404400) and Cervical Cancer "Medical Artificial Intelligence" (Grant No. WH10022021004). Yu Jiang is the primary corresponding author for the paper.

Author contributions

Conceptualization and project administration: Z.Y., Y.Q., G.G., and Y.J. Methodology: Z.Y., Y.Q., G.G., and Y.C. Visualization: Z.Y. Writing—original draft: Z.Y., Y.C., X.H., M.C., and W.W. Writing—review & editing: Z.Y., Y.C., X.H., M.C., W.W., H.Z., K.H., J.W., Q.L., Y.Z., Y.Q., and Y.J. Supervision: Y.Q., G.G., and Y.J.

Competing interests

The authors declare no competing interests.

Disclosing the use of artificial intelligence

An AI-assisted language tool (Gemini, Google) was used only for minor linguistic polishing. All study design, conceptualization, methodology, visualization, interpretation, and scientific writing were performed entirely by the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-03155-7.

Correspondence and requests for materials should be addressed to Zichen Ye, Yimin Qu, Guanghan Gao or Yu Jiang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026