Evaluating clinical AI before it touches patients.

A working framework for physicians and health-system leaders: what reported metrics mean and hide, what regulatory status actually establishes, what the 2026 benchmark and trial landscape can and cannot tell you, the failure modes of vendor demonstrations, and a one-page checklist you can bring to the next meeting. Every claim carries a numbered source. As of August 2026.

71%

of US non-federal acute care hospitals used predictive AI integrated with the EHR in 2024 [35]

6%

of 516 studies of imaging AI validated the model outside the institution that built it [2]

1,016

FDA authorizations reviewed in a peer-reviewed taxonomy; none involved a large language model [11]

Executive summary

Clinical AI is no longer a research topic. The US Food and Drug Administration (FDA) stated in January 2025 that it had authorized more than 1,000 AI-enabled medical devices [9], and 71% of US non-federal acute care hospitals reported using predictive AI integrated with their electronic health record (EHR) in 2024 [35]. The evidence behind those tools has not kept pace with their spread: in one review of 516 studies of AI for diagnostic imaging, only 6% performed external validation [2], and a systematic review of machine-learning prediction models rated 87% of analyses at high risk of bias [3]. Regulation will not close that gap for you. A 510(k) clearance establishes substantial equivalence to an earlier device, and in a peer-reviewed census of machine-learning devices the FDA authorized in 2024, only 29.2% of public summaries reported both sensitivity and specificity [12]. The European Union's AI Act deadlines for high-risk health AI have moved to December 2027 and August 2028 [20], and benchmark scores for large language models, however carefully built, measure answer quality on curated tasks rather than patient outcomes [25]. So the deciding evaluation happens where it always has: in your organization, against your patients, before go-live. This paper gives a physician or health-system leader the working tools for that evaluation — what reported metrics mean and hide, what regulatory status actually establishes, what the 2026 benchmark and trial landscape can and cannot tell you, the failure modes of vendor demonstrations, and a one-page checklist you can bring to the next meeting.

"Every layer of that infrastructure — the clearance, the leaderboard, the conformity assessment — evaluates the product in general. Only you can evaluate it for your patients."

Contents

1 The gap between adoption and evidence

71%

of US non-federal acute care hospitals used predictive AI integrated with the EHR in 2024.

6%

of 516 studies of imaging AI validated the model on data from outside the institution that built it.

Sources. ASTP/ONC Data Brief No. 80, September 2025, N = 2,253, response rate 51% [35]. Kim et al., *Korean Journal of Radiology*, 2019 [2].

Two numbers describe the state of clinical AI in 2026. The first is adoption: 71% of US hospitals used predictive AI integrated with the EHR in 2024, per a federal survey of 2,253 non-federal acute care hospitals [35]. The second is evidence: when researchers examined 516 studies of AI algorithms for diagnostic analysis of medical images, 31 of them — 6% — had been validated on data from outside the institution that built the model [2]. Not one of those 31 combined a diagnostic cohort design, multiple institutions, and prospective data collection.

The gap between those numbers is where patient harm lives. The best-documented example is a sepsis prediction model that shipped inside a widely deployed EHR. When University of Michigan researchers externally validated it on 38,455 hospitalizations, the model's area under the receiver operating characteristic curve (AUROC, explained below) was 0.63 — barely better than chance for a clinical tool — and at the alert threshold the study evaluated (a score of 6 or higher) it failed to identify 1,709 of the 2,552 patients who actually had sepsis (67%), while generating alerts on 18% of all hospitalizations [1]. Hundreds of hospitals were running it. The model had not been hidden; it had simply not been independently checked against the patients it was scoring.

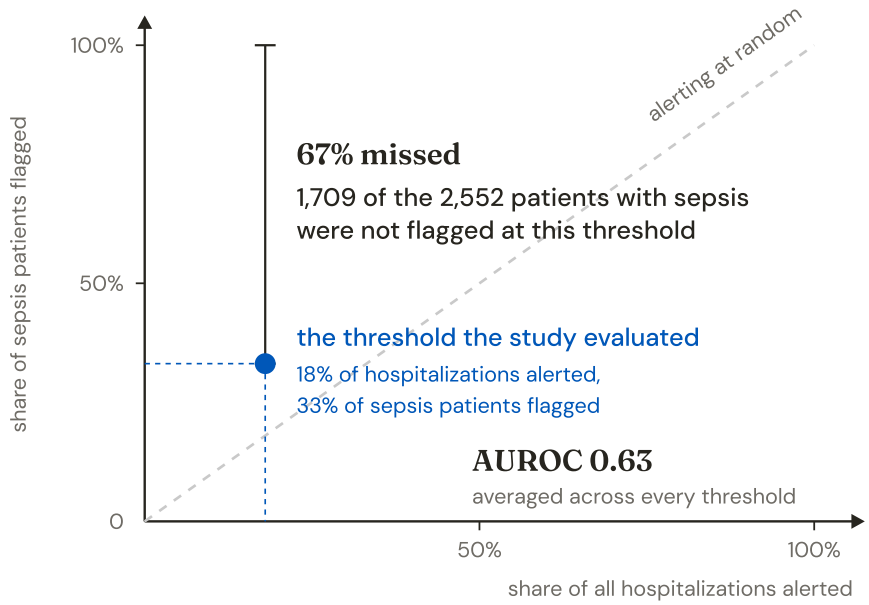
0.63

external AUROC of that model, measured on 38,455 hospitalizations at a hospital that took no part in building it.

67%

of the patients who actually had sepsis were not identified at the threshold the study evaluated.

Source. Wong et al., *JAMA Internal Medicine*, 2021 [1]. Alerts were generated on 18% of all hospitalizations.



the curve is a summary; the ward lives at the dot

Figure 1 — AUROC summarises how a model ranks patients across every threshold. A hospital does not deploy a curve; it deploys one point on it, and that point fixes sensitivity, specificity, and the daily alert volume clinicians absorb.

Drawn from the external validation reported in Wong et al., 2021 [1]. Only the marked point is an observed result.

That pattern is structural, not anecdotal. A BMJ systematic review of 152 machine-learning prediction-model studies rated 87% of its analyses (95% confidence interval 81% to 91%) at high risk of bias; 56% of models were developed with an inadequate number of outcome events for the number of candidate predictors, and 39% assessed overfitting improperly [3]. Another BMJ review found that 61 of 81 non-randomised studies comparing deep learning with clinicians claimed performance comparable to or better than doctors — while only nine of the 81 were prospective, and data and code were unavailable in 95% and 93% of studies respectively [4].

None of this means the field is empty of good tools. It means the burden of proof sits with the buyer, and the rest of this paper is about how to carry it.

2 What the number on the slide means

Most vendor decks lead with a single number. Reading it well takes four moves.

AUROC is a summary, and deployment happens at a point. The area under the receiver operating characteristic curve summarises how well a model ranks patients — whether it scores the patient who will deteriorate above the patient who will not — averaged across every possible alerting threshold, including thresholds nobody would run. An AUROC of 1.0 is perfect ranking; 0.5 is a coin flip. Your hospital will not deploy a curve. It will deploy one operating point: a threshold that fixes the sensitivity (the share of true cases flagged), the specificity (the share of non-cases correctly left alone), and, with them, the daily alert volume your clinicians absorb. The TRIPOD+AI reporting statement — the 27-item standard for studies of clinical prediction models, which supersedes the 2015 original [30] — expects performance to be reported in exactly these terms. Ask for the table of sensitivity and specificity at several thresholds, the threshold the vendor recommends, and the alert rate that follows from it. A vendor who can only produce the AUROC has told you the model ranks; they have not told you what it will do on a Tuesday.

"A vendor who can only produce the AUROC has told you the model ranks; they have not told you what it will do on a Tuesday."

Positive predictive value depends on your prevalence, and it is arithmetic you can do in the meeting. Positive predictive value (PPV) is the share of alerts that are right. It is not a property of the model alone; it moves with how common the condition is in your population. A tool with 90% sensitivity and 90% specificity, applied where 2% of patients have the condition, produces alerts that are right about 16% of the time — roughly five false alarms for every true one. That is the same model that posted "90% accuracy" on the slide, and the calculation takes one minute with your own base rate. The sepsis example above is what this looks like at scale: high alert volume, low yield, and a workload that clinicians eventually learn to ignore [1].

Discrimination is not the same property as calibration. Discrimination (AUROC) asks whether the model ranks sicker patients higher. Calibration asks whether a predicted 20% risk corresponds to a real 20% event rate. The two come apart, and calibration is the one that quietly fails: methodologists have called it the Achilles heel of predictive analytics, and note that it receives little attention in published models [6].

In a nine-year Veterans Affairs study of acute kidney injury models, discrimination held steady across the full period while calibration deteriorated and every model increasingly over-predicted risk [7]. If a tool feeds a risk number into a clinical conversation — a consent discussion, a triage decision, a threshold for treatment — ask for the calibration plot, its slope and intercept, and the population and date it was measured on.

Aggregate performance can conceal a failing subgroup. A model can hold its headline numbers while performing poorly for a group the topline averages over. The canonical case is the commercial risk score, applied to millions of US patients, that used healthcare cost as its proxy for healthcare need; because less money is spent on Black patients at the same level of illness, Black patients were systematically under-referred, and fixing the bias would have raised the share of Black patients receiving extra help from 17.7% to 46.5% [5]. Ask for performance broken down by the subgroups in your own catchment — age bands, sex, language, skin tone where imaging is involved — with the sample size in each cell, and an honest statement of which cells were too small to say anything about.

17.7% to 46.5%
The share of Black patients who would have been referred for additional help had the risk score not used cost as its proxy for need.

Obermeyer et al., *Science*, 2019 [5].

AIMOCS keeps plain-language explainers on each of these — AUROC, calibration, subgroup audits — in the article library at aimocs.com/learn.

3 Whose patients the evidence is about

The second question is not how good the number is. It is where the number came from.

Internal validation measures memory; external validation measures travel. Performance on held-out data from the same institution that trained the model tells you the model learned that institution — its scanners, its coding habits, its case mix. External validation, on data from sites that took no part in development, is the only evidence that a model's performance travels. It is rare enough that you should treat its absence as the default: 6% in the imaging review above [2]. When external results exist, ask for them per site, with the years the data was collected. A single pooled figure across "multiple health systems" can hide the site where the model did not work.

Dataset shift is the reason yesterday's validation expires. Models are trained on a snapshot of clinical practice, and practice moves: new scanners, new coding incentives, a new sepsis definition, a pandemic, a different payer mix. The New England Journal of Medicine essay on this problem — written partly by authors who watched their own hospital's models misfire when COVID-19 changed the inputs — gives it the name clinicians should know: dataset shift [8]. The practical consequences are two. First, evidence has a calendar: performance measured on 2019 data is a claim about 2019. Second, deployment creates a duty of monitoring, because the population under the model keeps moving after go-live (section 8).

The reporting and appraisal standards exist; make the vendor meet them. You do not need to invent your own quality bar. TRIPOD+AI defines what a complete report of a prediction model contains [30]. PROBAST+AI gives an appraisal tool — 16 signalling questions for how a model was developed and 18 for how it was evaluated — and names healthcare professionals verifying a model before purchase among its intended users [31]. DECIDE-AI defines what an early live clinical evaluation should report [32], and CONSORT-AI adds 14 AI-specific items to the standard for randomised trial reports [33]. FUTURE-AI, a consensus of 117 experts from 50 countries, sets out six principles and 30 recommendations spanning the whole lifecycle [29]. None of these was written as a procurement instrument, which is why AIMOCS reconciled them into a buyer's checklist — with each item traced to its source framework — in the clinical AI evaluation kit at aimocs.com/kit. The one-page version is section 7 of this paper.

The five standards. TRIPOD+AI, reporting [30]. PROBAST+AI, appraisal [31]. DECIDE-AI, early live evaluation [32]. CONSORT-AI, randomised trials [33]. FUTURE-AI, lifecycle principles [29].

Source. Singh et al., *npj Digital Medicine*, 2025 [11].

4 What FDA authorization actually establishes

"FDA-approved AI" is doing a lot of work in sales conversations. Here is what the record supports.

1,016

FDA authorizations reviewed in a peer-reviewed taxonomy of how AI is used in authorized medical devices. Quantitative image analysis was still the most common function; none of the 1,016 involved a large language model at the time of review.

The scale is real. The FDA said in January 2025 that it had authorized more than 1,000 AI-enabled devices through established premarket pathways [9], and it maintains a public list of them [10]. A peer-reviewed taxonomy that reviewed 1,016 of those authorizations found quantitative image analysis still the most common function [11]. The concentration is extreme: of the 168 machine-learning-enabled Class II devices authorized in 2024, 74.4% were in radiology, with cardiovascular next at 6.5% [12].

The pathway defines the proof. Nearly everything comes through the 510(k) pathway — 94.6% of the 2024 cohort [12] — which establishes that a device is substantially equivalent to a legally marketed predicate device [13]. That is a real review, and it is not a finding that the device improves outcomes, or that it performs at your hospital, on your scanners, at your prevalence. De Novo and premarket approval (PMA) routes carry different evidence, and many clinical AI tools — including most administrative and some decision-support software — are not FDA-reviewed devices at all. The first question about any regulatory claim is therefore: which pathway, which submission number, and what did it test. AIMOCS publishes a guide to reading a clearance summary at aimocs.com/guides/how-to-read-an-fda-clearance-summary.

Public transparency is thinner than the count suggests. In the 2024 cohort, only 29.2% of authorization summaries reported both sensitivity and specificity, and 15.5% provided demographic data on the populations tested [12]. An authorization, in other words, usually does not hand you the numbers sections 2 and 3 say you need. You have to ask the vendor, and the checklist assumes you will.

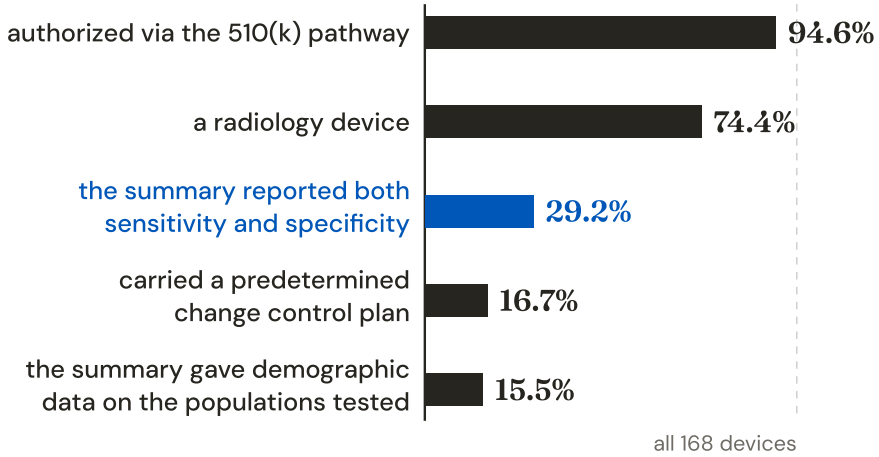


Figure 2 — What the public record contained for the 168 machine-learning-enabled Class II devices the FDA authorized in 2024. The blue bar is the one a buyer needs, and usually does not get.

Almarie et al., *Biomedicines*, 2025 [12].

an authorization is a permission, not a data sheet

Authorized devices can change after clearance. The FDA finalized guidance in December 2024 for Predetermined Change Control Plans (PCCPs): a mechanism by which a manufacturer can pre-specify future modifications — including model retraining — and ship them without a new submission, provided the plan includes a description of the modifications, a modification protocol, and an impact assessment [14]. PCCPs are legitimate and appeared in 16.7% of 2024 summaries [12]. For a buyer the implication is direct: ask whether the product has a PCCP, what it permits the vendor to change without telling you, and how a change reaches your site. Separately, the FDA issued a comprehensive draft guidance in January 2025 covering AI-enabled devices across their total product lifecycle [9]; as of August 2026 it remains draft, so its recommendations — including local site-specific acceptance testing — are not yet binding [15].

Generative AI has reached the pathway, barely. The taxonomy of 1,016 authorizations found that although over 100 devices use AI for data generation, none involved large language models (LLMs) [11]. The frontier moved in December 2025, when the FDA cleared UpDoc (510(k) K253281), a prescription insulin-management device whose patient interface is an LLM; the conversational layer collects and structures information while a deterministic, provider-configured algorithm computes every dose, and the clearance carries an authorized PCCP [16, 37]. The architecture is the lesson: what the agency accepted was an LLM wrapped around protocolized logic, with the clinical decision kept out of the generative component. For the many LLM products marketed to clinicians with no device authorization at all, evaluation falls entirely on you (section 6). AIMOCS tracks the device-list statistics, updated quarterly, at aimocs.com/guides/fda-ai-enabled-device-list-statistics-tracker.

The architecture is the lesson. An LLM wrapped around protocolized logic, with the clinical decision kept out of the generative component [16, 37].

One more US lever worth knowing: under the HTI-1 final rule on certification-program updates and algorithm transparency, published January 2024, predictive decision support supplied through certified health IT must make defined source-attribute information — intended use, development data, performance — available to users [17]. If the tool runs inside your certified EHR, you do not have to negotiate for that disclosure; you can simply require it.

5 Europe: the clock was reset, not stopped

The EU AI Act — Regulation (EU) 2024/1689 — classifies most clinical AI as high-risk, either as a standalone Annex III system or because it is a medical device under existing EU device law [18]. In 2026 its calendar changed, and buyers should know both what moved and what did not.

The Digital Omnibus on AI, Regulation (EU) 2026/1744, was adopted on 8 July 2026, published in the Official Journal on 24 July, and entered into force on 27 July 2026 [19]. Per the European Commission's own timeline, the high-risk obligations now apply from 2 December 2027 for Annex III systems, and from 2 August 2028 for AI embedded in regulated products — the route medical devices take [20]. What did not move: the Act became generally applicable on 2 August 2026, and the Article 50 transparency duties began on schedule — people must be told when they are interacting with an AI system, and AI-generated content must be identifiable [20]. A hospital deploying a patient-facing chatbot in the EU is already inside those duties.

The European calendar, after the Omnibus

Date	What applies to a hospital deploying clinical AI in the EU
Today	CE marking under the Medical Device Regulation, for AI that is a medical device. The AI Act adds obligations on top; it does not suspend device law in the meantime [18, 20].
2 August 2026	The AI Act became generally applicable, and the Article 50 transparency duties began on schedule: people must be told when they are interacting with an AI system, and AI-generated content must be identifiable [20].
2 December 2027	High-risk obligations apply to standalone Annex III systems [20].
2 August 2028	High-risk obligations apply to AI embedded in regulated products — the route medical devices take [20].

Regulation (EU) 2024/1689 [18]; Regulation (EU) 2026/1744, in force 27 July 2026 [19]; European Commission application timeline [20]. A digest of section 5; the deferral changes compliance dates, not the clinical evidence a buyer should demand.

Two practical notes. First, deferral of the high-risk dates is relief for compliance teams, and it changes nothing about the clinical evidence a buyer should demand; the Act was never going to externally validate a model on your population. Second, AI that is a medical device still needs its CE marking under the Medical Device Regulation today; the 2028 date adds AI Act obligations on top of that, and does not suspend device law in the meantime [18, 20]. AIMOCS maintains a living timeline at aimocs.com/guides/eu-ai-act-for-healthcare-living-timeline. For global context beyond the US and EU, the World Health Organization's guidance on large multi-modal models sets out governance recommendations for exactly the generative systems now entering care [36].

6 Benchmarks, and the evidence hierarchy for LLMs

Large language models broke the old evaluation playbook. A scribe or chatbot has no single AUROC; its output is text, judged case by case. The field's answer in 2025 and 2026 has been structured benchmarks, and the three worth knowing are these.

MedHELM (Stanford, published in Nature Medicine 2026) is a clinician-validated taxonomy of 121 medical tasks across five categories — decision support, note generation, patient communication, research, administration — with a benchmark suite covering all 22 subcategories, scored partly by juries of AI evaluators against expert-defined criteria [21]. Its public leaderboard turns over quickly: on the 14 May 2026 refresh, Gemini 3.1 Pro (Preview) led with a mean win rate of 0.652 [22]. **HealthBench** (OpenAI, 2025) grades 5,000 multi-turn health conversations against 48,562 physician-written rubric criteria authored by 262 physicians [23]. **HealthBench Professional** (OpenAI, 2026) narrows to the clinician side: 525 physician-authored tasks across care consult, documentation, and research, selected from 15,079 candidates with difficult cases enriched 3.5-fold, about a third of them deliberately adversarial, each rubric adjudicated by three or more physicians [24].

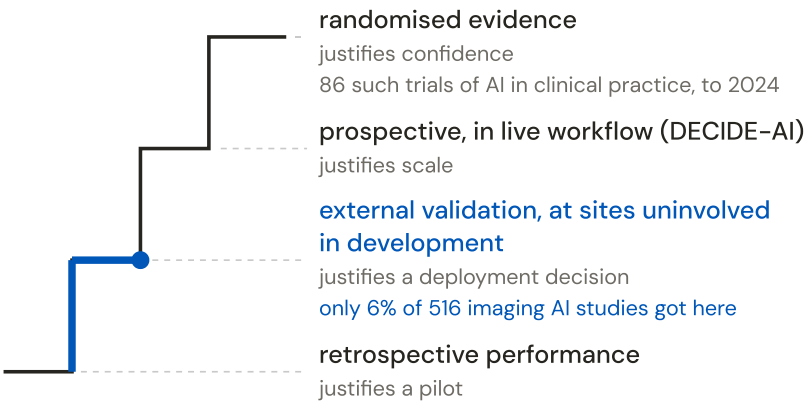
The most quoted result of 2026 comes from that last benchmark: the deployed ChatGPT for Clinicians system scored 59.0, above base GPT-5.4 at 48.1 and above physician-written responses at 43.7 — physicians who were specialty-matched and given unlimited time and web access [24]. Read carefully, the same paper contains its own caveats: the system's clearest advantage was on the hardest deliberately adversarial cases (55.8 versus 30.0 for physicians), while on the difficult good-faith slice — the closest analogue to hard, ordinary clinical questions — it was statistically tied with the physicians [24]. And the whole comparison is a vendor-reported result on a vendor-built benchmark: the tasks and rubrics are published, but the grading implementation behind the reported scores is internal, so independent re-scoring means rebuilding the grader [24].

5%
of 519 published evaluations of large language models in healthcare used real patient-care data. Calibration and uncertainty were measured in 1.2%.

Bedi et al., *JAMA*, 2025 [25].

What should a score like that change in your decisions? It should change which hypotheses you take seriously. It should not change what counts as clinical evidence, for three reasons grounded in the literature. First, benchmarks measure answer quality on curated conversations; a JAMA systematic review of 519 LLM-in-healthcare evaluations found only 5% used real patient-care data, 44.5% tested licensing-exam-style questions, and 95.4% scored accuracy while calibration and uncertainty were measured in 1.2% [25]. Second, benchmark tasks are static, while deployment is a workflow: interruptions, incomplete charts, users who trust or ignore the tool in patterned ways. Third, the leaderboard itself is unstable — the models that led the 2025 evaluation no longer lead it a year later [21, 22] — which is a reason to buy monitoring, and no reason to buy a ranking.

The hierarchy that should govern purchasing is the same one medicine already uses. Retrospective performance justifies a pilot. External validation justifies taking a deployment decision seriously. Prospective evaluation in live workflow — the DECIDE-AI stage [32] — justifies scale. Randomised evidence justifies confidence, and it remains scarce: a 2024 scoping review found 86 randomised controlled trials of AI in clinical practice, 81% reporting positive primary endpoints but concentrated in single centres with little demographic reporting [26].



most of the field never leaves the first step

Figure 3 — The evidence hierarchy that should govern purchasing. Each step licenses a different decision, and the blue step — validation at a site that took no part in building the model — is the one most tools never reach.

Kim et al., 2019 [2]; Vasey et al., DECIDE-AI, 2022 [32]; Han et al., *The Lancet Digital Health*, 2024 [26].

The live evidence runs in both directions, even within cancer screening. In Sweden's MASAI trial, 105,934 women were randomised and AI-supported mammography screening raised sensitivity to 80.5% from 73.8% at identical 98.5% specificity, with a non-inferior interval-cancer rate and reduced reader workload [27]. In Poland, an observational study across four endoscopy centres found clinicians' unassisted adenoma detection rate fell from 28.4% to 22.4% in the months after routine AI assistance began — a 6.0-point absolute drop, and a multicentre signal that a tool can change the clinician who uses it, and not only for the better [28]. Both results are tracked, alongside their caveats, at aimocs.com/statistics/clinical-ai-trial-results-tracker and the benchmark numbers at aimocs.com/statistics/llm-medical-benchmark-results-tracker.

7 The vendor demo, and the one-page checklist

Demos are optimized environments: curated cases, the vendor's hardware, the presenter's pacing. The failure modes repeat, and they are recognisable.

The enriched sample. Accuracy quoted from a test set with far more positives than your clinic sees. The tell: no prevalence stated. The fix: the PPV arithmetic from section 2, done on your base rate.

The pooled external validation. One flattering figure across "multiple health systems." The fix: per-site results, with years and sample sizes.

"FDA-approved" as one word. The fix: pathway, submission number, and the indications-for-use statement read verbatim — it names the patients, the setting, and the role of the output, and it is routinely narrower than the sales conversation [13].

The configurable threshold. "You can tune it to your needs" means the operating point, alert burden, and PPV are undetermined — the evaluation has been deferred to you without being named as such.

"The model improves continuously." That sentence describes either a PCCP the vendor should show you or an ungoverned change process [14]. Both are answers you need before signing.

The missing churn. Reference customers are selected. Ask for three customers who stopped using the product, and why. A mature vendor knows exactly why they lose sites.

The demo that never touches your data. The strongest single ask in procurement: run the tool, or its vendor-run equivalent, on a retrospective sample of your own cases, with the pass mark agreed before the result is known.

Seven failure modes. Each one is recognisable on sight, and each has a fix you can ask for before the meeting ends.

The checklist below compresses the full AIMOCS evaluation kit — which reconciles the published frameworks item by item, with sources — into one page. Anything unchecked is not automatically disqualifying; it is a finding to be explained in writing before signature.

The one-page checklist

Anything unchecked is not automatically disqualifying; it is a finding to be explained in writing before signature.

Evaluating clinical AI before it touches patients — AIMOCS, August 2026. Bracketed numerals point to the sources on pages 17 to 19.

A. Is the evidence about your patients?

- ☐ Indications-for-use statement in hand, verbatim, from the clearance or labelling — and your intended use falls inside it.
- ☐ Development population described (age, sex, case mix) and the differences from yours written down.
- ☐ Prevalence in the study data stated; your own prevalence known; expected PPV and alert burden computed from both.
- ☐ Calendar years, sites, and equipment/EHR versions of the data stated.

B. How was it tested?

- ☐ External validation exists, at sites uninvolved in development, reported per site [2, 30].
- ☐ Performance given at a named operating point (sensitivity, specificity, alert rate), and not AUROC alone [30].
- ☐ Calibration reported, with a plot, and dated [6, 7].
- ☐ Subgroup results with denominators, covering the groups in your catchment [5, 29].
- ☐ Any prospective or live-workflow evaluation, reported to DECIDE-AI or CONSORT-AI standards, if the tool drives decisions [32, 33].

C. What is its regulatory footing?

- ☐ Pathway and submission number stated, or an explicit statement that the tool is not a regulated device and why [13].
- ☐ PCCP obtained if one exists: what it permits, and how changes reach you [14].
- ☐ For EHR-embedded predictive decision support in the US: source attributes supplied under HTI-1 [17].
- ☐ For EU use: CE marking under device law now; AI Act high-risk obligations calendared for 2027/2028; Article 50 transparency duties met today [18, 20].

D. Will it survive contact with your workflow?

- ☐ Local validation on your own retrospective data, with a pre-agreed pass mark, before go-live [15].
- ☐ Expected alerts per hundred patients per day, and who absorbs them.
- ☐ Out-of-distribution behaviour described: what the system does when it does not know.
- ☐ Named clinical owner, and a training plan for actual users.

E. What happens after the signature?

- ☐ Monitoring plan: metrics (discrimination, calibration, subgroups), cadence, thresholds, a named owner on each side, and a dashboard you can see [7, 34].
- ☐ Version-change policy: notice, release notes, the right to decline or roll back.
- ☐ Data terms: what leaves your environment, retention, and whether any of it trains the vendor's models.
- ☐ Exit terms: what you keep — outputs in the record, audit trails, monitoring history — if you turn it off.
- ☐ Stop conditions, decided now and written down.

Compressed from the AIMOCS clinical AI evaluation kit, which reconciles TRIPOD+AI, PROBAST+AI, DECIDE-AI, CONSORT-AI and FUTURE-AI item by item, with sources — aimocs.com/kit.

8 After go-live: evaluation becomes governance

The evaluation does not end at signature, because the model's world does not stop moving: calibration drifts [7], populations shift [8], vendors ship changes [14]. The structures for living with that are now reasonably well described. The Joint Commission and the Coalition for Health AI (CHAI) published joint guidance in 2025 naming seven elements of responsible AI use — governance structures, privacy and transparency, data security, ongoing quality monitoring, blinded safety-event reporting, risk and bias assessment, and education — while describing itself as an initial, high-level document [34]. Practice is ahead of the guidance in one respect: 66% of US hospitals using predictive AI already name a specific committee or task force as accountable for evaluating their models, and 82% report evaluating for accuracy and 74% for bias [35].

66%
of US hospitals using predictive AI name a specific committee or task force as accountable for evaluating their models. 82% evaluate for accuracy; 74% for bias.

ASTP/ONC Data Brief No. 80, 2025 [35].

The committee is the right instrument, with one caveat: a body that can review but cannot restrict or retire a tool is a reading group. The tests of a real governance function are concrete. Approval thresholds set at go-live, in writing, by the committee rather than the vendor. A standing rule that a shipped version change triggers re-confirmation of the local performance check. A route from AI-related incidents into the existing patient-safety system rather than a new, slower one. And a fixed re-approval clock, so that nothing stays live by inertia. A full starter charter — seats, veto rights, risk tiers, escalation triggers, and a fourteen-field intake form — is published as part of the AIMOCS evaluation kit aimocs.com/kit, with the committee playbook at aimocs.com/guides/hospital-ai-governance-committee-playbook.

"A body that can review but cannot restrict or retire a tool is a reading group."

The honest summary of 2026 is this. The tools are real, the strongest of them now carry randomised evidence [27], and the regulatory and benchmark infrastructure is maturing fast. And still, every layer of that infrastructure — the clearance, the leaderboard, the conformity assessment — evaluates the product in general. Only you can evaluate it for your patients. The organizations that do well over the next five years will be the ones that could tell the difference between evidence and a slide, and that had written down, before go-live, what would make them turn the thing off.

Sources

Every source was retrieved and checked on 17 August 2026. Journal sources are cited by DOI; regulatory sources by the issuing body's own page.

1 Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine.* 2021;181(8):1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626 — Epic sepsis model: external AUROC 0.63; 1,709 of 2,552 sepsis patients (67%) missed at the evaluated threshold (score of 6 or higher); alerts on 18% of 38,455 hospitalizations.

2 Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images. *Korean Journal of Radiology.* 2019;20(3):405–410. https://doi.org/10.3348/kjr.2019.0025 — of 516 studies, 6% (31) performed external validation; none combined cohort design, multiple institutions, and prospective collection.

3 Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ.* 2021;375:n2281. https://doi.org/10.1136/bmj.n2281 — 148 of 171 analyses (87%, 95% CI 81%–91%) at high risk of bias; 56% inadequate events per predictor; 39% improper overfitting assessment.

4 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ.* 2020;368:m689. https://doi.org/10.1136/bmj.m689 — 61 of 81 non-randomised studies claimed performance at least comparable to clinicians; 9 prospective; data and code unavailable in 95% and 93%.

5 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–453. https://doi.org/10.1126/science.aax2342 — cost-as-proxy bias; remedying it would raise Black patients receiving additional help from 17.7% to 46.5%.

6 Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine.* 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7 — calibration receives little attention and poorly calibrated algorithms can be misleading and harmful.

7 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association.* 2017;24(6):1052–1061. https://doi.org/10.1093/jamia/ocx030 — over nine years of validation, discrimination was maintained while calibration declined and models increasingly over-predicted risk.

8 Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine.* 2021;385(3):283–286. https://doi.org/10.1056/NEJMc2104626 — the clinician-facing account of dataset shift and why deployed models decay.

9 US Food and Drug Administration. FDA Issues Comprehensive Draft Guidance for Developers of Artificial Intelligence-Enabled Medical Devices. Press announcement, 6 January 2025. https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-guidance-developers-artificial-intelligence-enabled-medical-devices — "more than 1,000 AI-enabled devices" authorized; announces the total-product-lifecycle draft guidance.

10 US Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices (device list). https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices — the live list; supports the existence and scope of the public registry.

11 Singh R, Bapna M, Diab AR, Ruiz ES, Lotter W. How AI is used in FDA-authorized medical devices: a taxonomy across 1,016 authorizations. *npj Digital Medicine.* 2025;8:388. https://doi.org/10.1038/s41746-025-01800-1 — 1,016 authorizations reviewed; none involved LLMs at the time of review.

12 Almarie B, Gonzalez-Gonzalez LF, dos Santos Barbosa LA, et al. Machine Learning-Enabled Medical Devices Authorized by the US Food and Drug Administration in 2024. *Biomedicines.* 2025;13(12):3005. https://doi.org/10.3390/biomedicines13123005 — 168 Class II devices in 2024; 94.6% via 510(k); 74.4% radiology; 29.2% reported both sensitivity and specificity; 15.5% demographic data; 16.7% PCCPs.

13 US Food and Drug Administration. Premarket Notification 510(k). https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k — the 510(k) standard: substantial equivalence to a legally marketed device.

14 US Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. Final guidance, originally issued 4 December 2024; reissued 18 August 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence — PCCP mechanism; description of modifications, modification protocol, impact assessment.

15 US Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft guidance, January 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing — total-product-lifecycle recommendations, including local acceptance testing; draft as of August 2026.

16 US Food and Drug Administration. 510(k) Premarket Notification K253281 (UpDoc). Decision 23 December 2025. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K253281 — the clearance record for the first prescription device with a patient-facing LLM interface; PCCP authorized.

17 US Department of Health and Human Services. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1), final rule. *Federal Register*, 9 January 2024. https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency — source-attribute transparency requirements for predictive decision support in certified health IT.

18 Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Official Journal of the European Union*, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj — the AI Act: high-risk classification and Article 50 transparency duties.

19 Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI), amending Regulation (EU) 2024/1689. *Official Journal*, 24 July 2026. https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng — the amending regulation; in force 27 July 2026.

20 European Commission. AI Act — application timeline (Shaping Europe's Digital Future). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai — post-Omnibus dates: Annex III high-risk from 2 December 2027; product-embedded (Annex I) from 2 August 2028; transparency enforced from 2 August 2026.

21 Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine.* 2026;32(3):943–951. https://doi.org/10.1038/s41591-025-04151-2 — 121 tasks, 22 subcategories, 5 categories; jury-based scoring.

22 Stanford Center for Research on Foundation Models. MedHELM leaderboard, v5.0.0, 14 May 2026. https://medhelm.org/ — current leaderboard standings; Gemini 3.1 Pro (Preview) mean win rate 0.652.

23 Arora RK, Wei J, Soskin Hicks R, et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv:2505.08775. 2025. https://arxiv.org/abs/2505.08775 — 5,000 conversations; 48,562 rubric criteria; 262 physicians.

24 Soskin Hicks R, Trofimov M, Lim D, et al. HealthBench Professional: Evaluating Large Language Models on Real Clinician Chats. arXiv:2604.27470. 2026. https://arxiv.org/abs/2604.27470 — 525 tasks from 15,079 candidates; scores 59.0 (deployed system), 48.1 (base GPT-5.4), 43.7 (physicians); adversarial-slice and good-faith-slice detail.

25 Bedi S, Liu Y, Orr-Ewing L, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA.* 2025;333(4):319–328. https://doi.org/10.1001/jama.2024.21700 — of 519 studies, 5% used real patient-care data; 44.5% exam-style questions; 95.4% accuracy-focused; 1.2% measured calibration and uncertainty.

26 Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health.* 2024;6(5):e367–e373. https://doi.org/10.1016/S2589-7500(24)00047-5 — 86 RCTs; 70 (81%) positive primary endpoints; single-centre predominance.

27 Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study. *The Lancet.* 2026;407(10527):505–514. https://doi.org/10.1016/S0140-6736(25)02464-X — 105,934 women randomised; sensitivity 80.5% vs 73.8%; specificity 98.5% in both groups; non-inferior interval-cancer rate.

28 Budzyń K, Romańczyk M, Kitala D, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology.* 2025;10(10):896–903. https://doi.org/10.1016/S2468-1253(25)00133-5 — unassisted adenoma detection rate fell from 28.4% to 22.4% after AI exposure (-6.0 points).

29 Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. https://doi.org/10.1136/bmj-2024-081554 — 117 experts, 50 countries; six principles; 30 best practices.

30 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378 — 27-item checklist; supersedes TRIPOD 2015.

31 Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505 — 16 development and 18 evaluation signalling questions across four domains; names healthcare professionals among users.

32 Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ.* 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904 — 17 AI-specific items (28 subitems) plus 10 generic, for early live evaluation.

33 Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine.* 2020;26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x — 14 new AI-specific trial-reporting items.

34 Joint Commission and Coalition for Health AI. Guidance on the Responsible Use of AI in Healthcare (RUAIH). 2025. https://assets.ctfassets.net/7s4afyr9pmov/5RX4XUbRg0l1JG0gwXGkTM/bea3c1949d6d8cdd8ec2ae2f82f737ad/JC-CHAI_RUAIH_Guidance.pdf — seven elements of responsible AI use; self-described initial, high-level document.

35 Chang W, Owusu-Mensah P, Everson J, Richwine C. Hospital Trends in the Use, Evaluation, and Governance of Predictive AI, 2023–2024. ASTP/ONC Data Brief No. 80, September 2025. https://www.healthit.gov/data/data-briefs/hospital-trends-use-evaluation-and-governance-predictive-ai-2023-2024 — 71% of hospitals used predictive AI; 82% evaluated accuracy, 74% bias; 66% named a specific committee; N=2,253, response rate 51%.

36 World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. https://www.who.int/publications/i/item/9789240084759 — WHO governance guidance for large multi-modal models in health.

37 Innolitics. First FDA-Cleared AI Agent and LLM-Enabled Device Confirmed. 2026. https://innolitics.com/articles/updoc-fda-cleared-ai-agent/ — regulatory analysis of K253281: LLM conversation layer around deterministic, provider-defined dosing logic.

Companion material

Each of the following is maintained and dated. Where a number in this paper is one that moves, the tracker carries the current value.

aimocs.com/kit
The clinical AI evaluation kit: vendor questions, the reconciled checklist with each item traced to its source framework, red flags, and a governance charter.

aimocs.com/guides/fda-ai-enabled-device-list-statistics-tracker
Device-list statistics, updated quarterly.

aimocs.com/statistics/llm-medical-benchmark-results-tracker
Benchmark standings for large language models on medical tasks.

aimocs.com/statistics/clinical-ai-trial-results-tracker
Trial results, with their caveats.

aimocs.com/guides/eu-ai-act-2-august-2026-what-applies
What the AI Act applies today, and what was deferred.

About this paper

Published by AIMOCS, a community for artificial intelligence in healthcare — physicians, healthcare executives, and researchers. Written and edited in house; the byline is the organization.

Every factual claim carries a numbered source that was retrieved and checked on 17 August 2026. Where a source is a living page, the retrieval date is the claim.

This paper is a working framework, not legal, regulatory, or clinical advice. Reproduce it for internal use within your organization, with attribution.

AIMOCS
aimocs.com

Evaluating clinical AI before it touches patients. August 2026.

Thirty-seven sources, checked 17 August 2026.