



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Clinical Evidence Requirements for AI Diagnostic Tools

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-02-08 · ai diagnostic tools · clinical evidence · medical device regulation · fda 510(k) · clinical validation · prospective studies · software as a medical device · eu mdr

Original article: <https://intuitionlabs.ai/articles/clinical-evidence-requirements-ai-diagnostics>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Regulatory Landscape for AI Diagnostic Tools
 4. Clinical Evidence Standards and Guidelines
 5. Evidence Generation and Study Design
 6. Case Studies and Examples
 7. Challenges and Considerations
 8. Discussion and Future Directions
 9. Conclusion
-

Executive Summary

Artificial intelligence (AI) is rapidly transforming medical diagnostics by analyzing complex clinical data—images, signals, and records—to assist or automate disease detection. A key barrier is the paucity of rigorous clinical evidence demonstrating safety, accuracy, and clinical utility. Experts note that “*little evidence exists so far for their effectiveness in practice*” for AI medical devices ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)), and even less is known about their safety or potential biases. In response, regulators and stakeholders worldwide now emphasize stringent evidence requirements tailored to AI’s unique attributes. Notably, the FDA and EU regulators require **robust clinical or performance evaluations using representative data** (www.osborneclarke.com). In the US, most AI tools have been cleared through the 510(k) pathway (Premarket Notification), which demands substantially *less* clinical evidence than a Premarket Approval (PMA) and instead relies on predicates ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). This “predicate creep” can leave safety and effectiveness inadequately established. In contrast, the EU’s Medical Device Regulation (MDR/IVDR) mandates *clinical evaluation* for diagnostic AI before CE marking, including trials or real-world data demonstrating intended performance (www.osborneclarke.com). The EU AI Act entered into force on 1 August 2024; according to the **Commission’s current timeline**, high-risk rules for qualifying Annex I regulated-product AI systems apply from 2 August 2028.

High-quality clinical evidence is needed at every stage of AI development. **Retrospective validation** on historical data is a necessary minimum, but by itself is *insufficient*. Prospective studies in intended clinical workflows are strongly preferred—ideally randomized trials or rigorous cohort studies that measure how AI use affects diagnoses, treatments, and patient outcomes. For example, the prospective IDx-DR trial enrolled 900 patients and reported sensitivity of 87.2% and specificity of 90.7% for more-than-mild diabetic retinopathy. FDA granted De Novo classification on 11 April 2018 for screening adults with diabetes who had not previously been diagnosed with diabetic retinopathy, using the Topcon NW400 ([FDA record](#); [indications and review summary](#)). This is a screening indication, not treatment or detection of other eye diseases. The Swedish MASAI trial evaluated interval cancer rates with a prespecified 20% non-inferiority margin ([primary trial report](#)). These findings underscore that **prospective clinical evaluation and statistical significance against pre-specified endpoints are critical**. Where trials are infeasible, well-designed real-world cohorts and registries—with transparent reporting of sensitivity, specificity, and calibration—can help build the evidence base.

Nevertheless, experts observe that **current evidence generation is fragmented** ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Many published studies rely solely on in silico or retrospective analysis, often with biased or non-representative data (www.sciencedirect.com) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). This leads to a “high risk of bias” identified in systematic reviews (www.sciencedirect.com). The lack of standardized reporting (as highlighted by CONSORT-AI and STARD-AI initiatives) makes it hard to compare studies or trust claims. Regulatory reviewers increasingly demand transparency: the FDA’s own **AI-Enabled Device List** stipulates that authorized devices have undergone appropriate clinical evaluation of safety and efficacy (www.fda.gov). The EU’s new guidance explicitly requires “*robust clinical evidence ... using data that is representative of the intended patient population*” (www.osborneclarke.com). In practice, companies are advised to document data quality, plan for bias mitigation, and engage in post-market surveillance to ensure algorithms remain safe as data drift and software updates occur (www.osborneclarke.com).

In this report, we analyze the current state and future prospects of clinical evidence for AI diagnostics. We review regulatory frameworks (US, EU, and other regions), detail the types and quality of data needed, and discuss study designs (from in silico trials to randomized controlled trials). We highlight multiple case examples—from autonomous retinal screening to stroke triage—to illustrate how evidence generation can succeed or fall short. We also examine expert opinions and published findings on necessary performance standards, validation strategies, and post-market monitoring. Ultimately, rigorous and standardized evidence generation will be essential to fully realize AI diagnostics’ promise of improved accuracy, efficiency, and patient outcomes.

Introduction and Background

Artificial intelligence (AI) in medicine promises to augment or even automate diagnosis by leveraging machine learning on large clinical datasets. An **AI medical diagnostic device (AIMDD)** uses algorithms to analyze multi-source data—such as symptoms, vital signs, medical images, and lab results—to determine disease type, stage, or severity, and to assist clinicians (www.sciencedirect.com). In practice, AI diagnostics often manifest as software-as-a-medical-device (SaMD). For example, convolutional neural networks in imaging (radiology, pathology) perform automated lesion detection or grading, while natural language processing can flag mentions of disease in clinical notes. AI tools are positioned as *second readers* or *prescreeners*, relieving workload and improving consistency by catching findings a human might miss (www.sciencedirect.com). Indeed, studies have shown that AI assistance can increase sensitivity and diagnostic consistency for clinicians (www.sciencedirect.com). Clinical implementations today include diabetic retinopathy screening (IDx-DR), stroke detection systems (e.g. Viz.ai), and tools for lung nodule triage, arrhythmia alerts, and more.

Barriers include the limited availability of high-quality, representative data for training; concerns over patient privacy and data security; and the “black-box” nature of many algorithms which hinders clinician trust (www.sciencedirect.com). Ethical and legal issues—such as algorithmic bias, liability for misdiagnosis, and equitable access—also complicate adoption. Crucially, the **lack of robust evidence of accuracy and safety** in real clinical settings undermines both regulators’ and clinicians’ confidence ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (www.sciencedirect.com). As one expert review observes, although AI diagnostics are widely anticipated to improve care processes, “*little evidence exists so far for their effectiveness in practice*”, raising questions about when, how much, and whom they truly help ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

Regulators, health systems, and professional groups are striving to address these gaps. International guidelines (e.g. SPIRIT-AI, STARD-AI, CONSORT-AI) have been proposed to standardize the design and reporting of AI studies ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). Regulatory agencies (FDA, EMA, NMPA, etc.) are updating rules and issuing guidance on AI-specific considerations (data quality, bias, updates). Concurrently, health technology assessment bodies like NICE have developed evidence frameworks to guide AI adoption ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (www.sciencedirect.com). Despite these efforts, the field is still maturing: as Weissman et al. note, “*the rate of investment has outpaced the rate of high-quality evidence production*” for AI systems ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). For AI diagnostics to be safe, effective, and trusted, a rigorous body of clinical evidence—on par with traditional medical devices—must be developed.

This report systematically examines **clinical evidence requirements** for AI diagnostic tools. We begin with an overview of regulatory landscapes globally, then delve into what constitutes appropriate evidence. We consider both statistical measures of performance (sensitivity, specificity, AUC, etc.) and clinical outcomes (impact on patient care). We discuss study designs—from retrospective validations to prospective trials—and challenges unique to AI (e.g. dataset shift, transparency). Multiple real-world case studies illustrate how evidence has been generated (or neglected) in practice. Throughout, we incorporate the latest data, expert analyses, and regulatory announcements to paint a comprehensive picture of current and emerging standards for evidence in AI diagnostics.

Regulatory Landscape for AI Diagnostic Tools

AI diagnostic software generally falls under medical device regulations worldwide. However, frameworks differ by jurisdiction:

United States (FDA)

In the U.S., medical-device status depends on the software function’s intended use and applicable statutory exclusions. FDA’s [January 29, 2026 final Clinical Decision Support Software guidance](#) explains the exclusion in FD&C Act section 520(o)(1)(E). A function must meet **all four criteria**: (1) it is not intended to acquire, process, or analyze a medical image, an IVD-device signal, or a pattern or signal from a signal-acquisition system; (2) it displays, analyzes, or prints patient medical information or other medical information; (3) it supports or provides recommendations to a health care professional about prevention, diagnosis, or treatment; and (4) it enables that professional to independently review the basis of the recommendations, so the professional is not intended to rely primarily on them for an individual patient’s diagnosis or treatment decision. Providing recommendations alone therefore does not establish device status. Image- or signal-analysis functions intended for a medical-device purpose do not qualify for this CDS exclusion.

FDA Pathways and Evidence Expectations

Once deemed a device, an AI diagnostic tool typically enters the FDA market via one of three pathways:

1. **Premarket Approval (PMA)**: The highest-risk category requiring substantial clinical evidence (usually from RCTs or trials) demonstrating safety and efficacy. Very few software products go PMA.

2. **Premarket Notification (510(k) or PMN):** The most common route for AI software. Here the manufacturer must show the device is “substantially equivalent” to a predicate device already legally marketed. Crucially, this requires *much less* clinical evidence than a PMA ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). In practice, most cleared AI tools rely on comparisons to older devices. Weissman et al. report that “*the PMN pathway that requires little clinical evidence if substantial equivalence to a predicate device is established may undermine efforts to establish safety and effectiveness for modern AI medical devices*” ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). They describe instances of “predicate creep,” where chains of predicates (some dating to the 1980s) vitiate meaningful evaluation ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). Indeed, since 2020 the FDA has maintained a public database of AI/ML devices showing that **the majority of FDA-authorized AI devices were cleared through 510(k)/PMN** ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
3. **De Novo Classification:** For novel, moderate-risk devices without a suitable predicate. This pathway can require clinical data but is less burdensome than PMA. Some AI systems (especially truly pioneering ones) may use De Novo.

The FDA explicitly reviews *safety and effectiveness evidence* in these applications. Its own AI-Enabled Devices list notes that included products “have met the FDA’s applicable premarket requirements, including a focused review of the device’s overall safety and effectiveness,” which entails evaluating study appropriateness for the device’s intended use (www.fda.gov). In practice, however, 510(k) clearances often involve limited clinical validation. Critics argue that many cleared AI tools have not undergone rigorous prospective evaluation ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). FDA’s 2022 guidance on Computer-Assisted Detection in Radiology, for example, still allows retrospective imaging data as primary evidence for 510(k) submissions, with only a recommendation to perform “clinical performance assessment” when possible ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).

The FDA is adapting to these concerns. In 2021 it released an *Action Plan for AI/ML-Based SaMD*, emphasizing “Good Machine Learning Practices” (GMLP) for development, real-world monitoring, and change management. The FDA also convened the AI/ML Consensus Conference to define best practices. Future regulatory updates will likely stress lifecycle oversight and outcome-based evaluation (as hinted in Weissman et al.: “priorities include ... a focus on patient health outcomes” ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/))). For now, however, the onus remains largely on developers to justify performance with available data.

European Union (MDR/IVDR and AI Act)

In the EU, AI diagnostic software is regulated under the **Medical Device Regulation (MDR)** or **In Vitro Diagnostic Regulation (IVDR)**, depending on claimed purpose. Since May 2021, the MDR demands a higher standard of clinical evidence than the previous MDD, with Notified Bodies required to review clinical data for all IIa+ devices. A central principle is *clinical evaluation*: manufacturers must demonstrate conformity by generating clinical data proving safety and performance per *State of the Art* and Claim requirements. For diagnostic software, this could involve clinical trials, prospective studies, or retrospective analyses, provided they use **representative patient data and real-world settings**.

Recent EU guidance (joint MDCG/AIB FAQ) clarifies specific expectations for *Medical Device Artificial Intelligence (MDAI)*. Key points include:

- Under MDR/IVDR, manufacturers must generate “robust clinical evidence through clinical or performance evaluation, using data that is representative of the intended patient population and use environment” (www.osborneclarke.com).
- The **AI Act** entered into force on 1 August 2024, with general application from 2 August 2026 and phased exceptions. As reviewed on 5 October 2026, the **European Commission’s timeline** places high-risk Annex I regulated-product rules at **2 August 2028** following the AI Omnibus. Existing MDR/IVDR duties remain applicable. Under **Article 6(1)**, the regulated-product route to high-risk status requires both an AI system that is a covered product or its safety component and a requirement for third-party conformity assessment under the relevant Annex I legislation. The medical-device risk class alone is not a complete AI Act classification test.
- The EU guidance notes that alongside MDR’s clinical evidence provisions, AI-specific rules call for comprehensive data governance and continuous surveillance to ensure post-market safety (www.osborneclarke.com).
- Notably, the EU rules do not fundamentally lower evidence standards for AI; they reinforce that even AI’s “self-learning” aspect must be controlled. The MDR requires that any significant software changes (including model updates) be documented, and high-risk AI likely will necessitate new conformity assessments after major updates (www.osborneclarke.com).

Clinical or performance evaluation should be matched to the device’s intended purpose and applicable MDR or IVDR requirements. For qualifying high-risk AI, AI Act obligations add a further layer under the application timeline above. Article 14 addresses effective human oversight with measures commensurate with risk, autonomy, and context of use; it does not adopt FDA’s CDS independent-review test (**AI Act, Article 14**).

Other Jurisdictions

- **United Kingdom:** In Great Britain, the UKCA route is based on the **UK Medical Devices Regulations 2002, as amended**, derived from the earlier EU directives. Qualifying CE-marked MDD/AIMDD devices may be placed on the market until certificate expiry or 30 June 2028, whichever is sooner; qualifying IVDD devices until certificate expiry or 30 June 2030; and MDR/IVDR-compliant devices until 30 June 2030, subject to the transition conditions. Northern Ireland applies the EU MDR/IVDR and requires CE marking, with UKNI also required when a UK notified body undertakes mandatory third-party conformity assessment (**MHRA market-access guidance**). NICE’s **Evidence Standards Framework** supports evaluation and purchasing decisions; meeting it is not regulatory approval or NICE endorsement.

The key takeaway is that **regulatory authorities now clearly require evidence appropriate to the risk and novelty of the AI diagnostic**. Notably, evidence requirements for AI—if not explicitly detailed in older frameworks—are being codified via guidelines, standards, and new legislation (e.g. the EU AI Act). For high-stakes diagnostics, both U.S. and EU regimes ultimately demand demonstration of clinical validity (accuracy) and often clinical usefulness. Critically, regulators emphasize that **data quality and representativeness are part of evidence quality**: it is not enough to report high AUC on a homogeneous dataset if the real-world patient mix differs (www.osborneclarke.com) (www.osborneclarke.com).

Clinical Evidence Standards and Guidelines

Given this regulatory landscape, what constitutes acceptable “clinical evidence” for AI diagnostics? While details vary, a consensus is emerging across academic and policy circles:

- **Performance Metrics as Minimum:** For diagnostic accuracy, key statistics include sensitivity, specificity, and area under the Receiver Operating Characteristic (ROC) curve (AUROC). These metrics should be reported with confidence intervals and on predefined endpoints. For example, pivotal AI trials often set superiority thresholds (as with IDx-DR) that must be met or exceeded. The **FDA has long emphasized sensitivity/specificity** (see its 2022 guidance on Computer-Assisted Detection) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/36111111/)). However, modern evaluators caution against relying solely on summary metrics: they advocate thorough calibration (to ensure predicted risks match observed rates) and decision curve analysis (to understand clinical impact) when possible.
- **Representative Datasets:** Demonstrating accuracy requires that test data reflect intended use. EU guidance explicitly calls for “data representative of the intended patient population and use environment” (www.osborneclarke.com). This echoes calls in the literature to avoid sampling bias. The scoping review by Xu et al. highlights that many AI studies suffer **spectrum bias** (e.g., evaluating mostly severe cases) and under-represent minorities (www.sciencedirect.com). Best practice is to include diverse, multi-center cohorts, and to perform **external validation** on entirely independent datasets (different hospitals, geographies, or time periods) (www.sciencedirect.com) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/36111111/)).
- **Clinical Utility and Outcomes:** Accuracy alone may not predict clinical benefit. Regulators increasingly ask: *does the AI improve patient care?* For example, MASAI used AI to support single or double reading by radiologists and reported higher screening sensitivity without increasing false positives ([primary trial report](#); [trial investigators' summary](#)). Such screening-performance findings should be distinguished from evidence measuring time to treatment. Randomized trials (or at least well-controlled prospective studies) can capture such utility. Level-of-evidence frameworks note that evidence on patient outcomes (e.g. morbidity, mortality, quality of life) can substantially strengthen a case for benefit.
- **Reporting Standards:** A growing suite of guidelines prescribes how to design, conduct, and report studies on AI diagnostics. Key examples include:
 - **CONSORT-AI** (Lancet Digital Health 2020): An extension of the CONSORT RCT checklist, requiring specification of AI interventions, algorithm versions, data pre-processing, and error analysis ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/36111111/)).
 - **DECIDE-AI** (BMJ 2022): Guidance for early-phase clinical evaluation of AI decision-support tools, emphasizing iterative assessment and transparency in algorithm changes ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/36111111/)).
 - **TRIPOD+AI** (BMJ, 16 April 2024): Updated reporting guidance for clinical prediction models developed or evaluated using regression or machine learning ([published statement](#)).

Guideline / Framework	Focus / Scope	Issued By / Year	Key Aspects
CONSORT-AI	Reporting of RCTs involving AI interventions	Lancet Digital Health (2020)	Extends CONSORT to include AI-specific details (algorithm, data drift, human-AI interaction) (pmc.ncbi.nlm.nih.gov).
DECIDE-AI	Early-stage evaluation of AI-driven decision support	BMJ (2022)	Checklist for designing and reporting first-in-human clinical studies of AI tools (pmc.ncbi.nlm.nih.gov).
TRIPOD+AI	Clinical prediction model development and evaluation using regression or machine learning	BMJ (16 April 2024)	Updated reporting checklist (statement).
NICE ESF	Evaluation of digital health technologies	NICE (March 2019; AI-focused update August 2022)	Standards supporting evaluation and purchasing decisions (framework).
FDA AI/ML Action Plan / GMLP	Quality system principles for AI medical devices	FDA (2021 – ongoing)	Emphasizes validation, real-world performance monitoring, predetermined plans for algorithm updates.
STARD-AI	AI diagnostic accuracy study reporting	Nature Medicine (15 September 2025)	Completed guideline with 18 new or modified items alongside STARD 2015 (guideline).

Table 1. Examples of selected guidelines and frameworks guiding the evaluation of AI diagnostic tools. Each provides criteria or checklists for evidence generation and reporting. Citations in the key aspects column indicate source discussions.

These guidelines reflect a consensus that **transparency and rigor are essential in AI studies**. As one review notes, initiatives like CONSORT-AI and DECIDE-AI are being developed “*in close collaboration with key stakeholders... to generate a consensus applicable across the community*” (pmc.ncbi.nlm.nih.gov). However, adoption is variable: many published AI studies still do not fully follow these guidelines, contributing to the heterogeneity and uncertainty noted by regulatory reviewers (pmc.ncbi.nlm.nih.gov) (www.sciencedirect.com).

Evidence Generation and Study Design

Producing credible clinical evidence for AI diagnostics involves multiple study designs. Ideally, a development pathway leads from preclinical validation to clinical validation and finally to outcome studies, each with appropriate rigor. We outline the common phases and their considerations:

In Silico and Retrospective Validation

Concept: Assessment of the AI model using existing datasets, often fully annotated, to estimate performance metrics. This is typically the first step, akin to technical validation.

- **Data Sources:** May include public datasets (e.g. publicly available imaging collections) or institutional data cohorts. For example, many early AI diagnostic papers use retrospective imaging archives for training and testing. However, Xu et al. caution that well-annotated public datasets remain limited

(www.sciencedirect.com).

- **Performance Metrics:** Sensitivity, specificity, positive predictive value, AUC, F1 score, etc. Cross-validation and holdout test sets are common. Calibration plots (predicted vs actual probability) and decision curve analyses are recommended for completeness.
- **Limitations:** Overfitting is a major risk if data are limited or not diverse (www.sciencedirect.com). Even a high AUC on internal validation does not guarantee real-world success. Key failure modes include lack of *external validity*: e.g., an algorithm trained on one hospital's CT machines may underperform on a different machine. There is also a trend of "data leakage" or bias: retrospective cohorts often exclude borderline cases or have spectrum bias (www.sciencedirect.com).
- **Regulatory View:** Such studies are necessary but only preliminary. The FDA generally expects at least a "clinical performance assessment" to accompany 510(k) for AI in imaging ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Some FDA guidances suggest retrospective data are acceptable for initial clearance, provided the databases are robust (e.g. Radiological Computer-Assisted Detection guidance).

Prospective Clinical Validation (In-Clinico Studies)

Concept: Evaluating the AI tool in a clinical setting going forward. This includes studies where the AI is integrated into practice (even if results are not used) or where samples are collected prospectively.

- **Non-Randomized Trials / Cohorts:** Many regulatory clearances rely on prospective observational studies. For example, the IDx-DR pivotal trial was a prospective single-arm study: primary care clinics enrolled patients without known DR, and the AI's decisions were compared to gold-standard grader assessments ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Similarly, a company might perform a prospective "silent mode" rollout where AI runs in the background while doctors render their usual diagnoses; discrepancies are later analyzed.
- **Randomized Controlled Trials (RCTs):** The gold standard for demonstrating impact on care. In the AI context, RCTs can randomize patients or providers to use-AI versus standard care. The completed MASAI trial is registered as NCT04838756 ([primary trial report](#)). RCTs control for confounders and can measure endpoints like stage at diagnosis or downstream treatment outcomes.
- **Trial Design Considerations:** Protocols must pre-specify primary outcomes (not just AUC) and analysis plans. The SPIRIT-AI and CONSORT-AI guidelines encourage detailing how the AI is locked (no changes during trial), how it will be deployed, and how clinician training is handled ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Endpoint examples include lesion detection rate, false referral rate, time to diagnosis, or resource utilization.
- **Regulatory Role of RCTs:** While not universally required, especially for lower-risk support tools, controlled study data can strengthen a clinical benefit claim. Reimbursement is a separate decision from FDA marketing authorization; NTAP is part of the [CMS inpatient payment program](#).

Real-World and Post-Market Studies

Even after regulatory clearance, evidence generation continues:

- **Post-Market Surveillance:** Regulators often require post-market surveillance plans. Consumer feedback, adverse event reports, and performance audits in deployment settings help detect issues (e.g. drift, rare errors). The AI Act will mandate additional post-market monitoring for high-risk AI medical devices.
- **Registries and Big Data:** Healthcare systems or consortia may set up registries tracking AI tool usage and outcomes. For instance, an AI for diabetic retinopathy might be monitored in a large diabetic screening program to measure real-world sensitivity and referral impact.
- **Technology Maintenance:** FDA's [final PCCP guidance](#) applies to AI-enabled device software functions that a manufacturer intends to modify over time, including automatic and manual model changes. A PCCP describes planned modifications, methods to develop, validate, and implement them, and an impact assessment. FDA reviews it within a marketing submission; it is not unrestricted permission for continued learning.
- **Reproducibility:** Replication in independent settings is part of evidence quality. The VALIDATE observational study compared hospitals using AI with hospitals without AI across 166 facilities. Its endpoint was **arrival-to-neurointerventionalist notification**, rather than door-to-puncture: median times were 50 versus 89.5 minutes, a 39.5-minute difference. The 2023 [manufacturer announcement](#) preceded a [2024 peer-reviewed report](#). This endpoint differs from the single-center puncture endpoint discussed below, and the observational design does not establish causation.

Statistical Considerations

Designing evidence generation also involves statistical rigor:

- **Sample Size and Power:** Diagnostic accuracy studies require enough cases (positive and negative) to estimate metrics precisely. Sample size calculations should be based on anticipated prevalence and desired confidence intervals. Underpowered studies risk misleadingly wide CIs.
- **Bias Control:** Randomization is the best control; if not used, careful cohort matching and adjustment (or case-control design) should be considered. Cross-validation must avoid data leakage.
- **Multiple Comparison and Overfitting:** When many models or cutoffs are tested, corrections or a holdout test set is needed to avoid optimistic bias.
- **Reporting Uncertainty:** All performance numbers should carry confidence intervals. Decisions should account for the inherent stochasticity of AI outcomes.
- **Health Economics:** Increasingly, evidence dossiers include cost-effectiveness analyses. For health systems to invest, they look for data on how AI tools change hospital stay, require fewer follow-ups, etc.

Case Studies and Examples

To ground these concepts, we examine several high-profile AI diagnostic tools as illustrative case studies. Each highlights specific evidence strategies and regulatory outcomes:

1. IDx-DR: Autonomous Diabetic Retinopathy Detection

Overview: FDA granted **De Novo classification** for IDx-DR on 11 April 2018 ([FDA record](#)). The Class II device is indicated for health care providers to automatically detect more-than-mild diabetic retinopathy in adults diagnosed with diabetes who have not previously been diagnosed with diabetic retinopathy, using the Topcon NW400. It is prescription-only, does not treat retinopathy, and is not intended to detect other eye diseases. Its outputs include a screening result or an insufficient-quality result; appropriate follow-up remains necessary ([FDA indications, limitations, and decision summary](#)).

Clinical Trial Evidence: The FDA De Novo review included a pivotal prospective study of **900 patients** with diabetes, none previously diagnosed with DR ([pmc.ncbi.nlm.nih.gov](#)). Patients underwent retinal imaging, and IDx-DR's output was compared to an expert reading center's grading using the ETDRS gold standard. The primary endpoint was to demonstrate sensitivity >85% and specificity >82.5% for detecting *more-than-mild* DR. The results were very positive: IDx-DR achieved 87.2% sensitivity (95% CI 81.8–91.2%) and 90.7% specificity (95% CI 88.3–92.7%), exceeding both prespecified thresholds ([pmc.ncbi.nlm.nih.gov](#)). Its imageability (proportion of participants with a disease-level AI output) was also very high at 96.1%. The trial was published in *NPJ Digital Medicine* ([pmc.ncbi.nlm.nih.gov](#)) following the 2018 De Novo authorization, and coverage noted that the 900-patient trial “beat the pre-specified endpoints for sensitivity and specificity” ([www.medtechdive.com](#)). The FDA decision summary describes the prospective clinical study and its performance analyses ([FDA review](#)).

Real-World Performance: Subsequent studies have examined IDx-DR in practice. A 2026 German study reported that when excluding ungradable images, the system achieved up to 94.4% sensitivity and 90.5% specificity for severe DR ([pmc.ncbi.nlm.nih.gov](#)). However, the study also highlighted imageability issues: when including non-gradable cases, sensitivity dropped. This reflects an important evidentiary point: the prospective trial had strict training of personnel on image capture, whereas real-world sites vary. Such post-market evidence underscores the need for **robust data acquisition protocols** as part of clinical utility demonstration.

Regulatory Impact: The [2018 De Novo grant](#) is distinct from PMA approval and 510(k) clearance. Reimbursement is also a separate process: CMS's CY 2021 OPPS final rule discusses IDx-DR under **CPT 92229**, effective 1 January 2021, and assigns the code to outpatient APC 5733 with status indicator S ([CMS final rule](#)). This is an outpatient payment provision, not an inpatient NTAP award or an FDA finding of payment eligibility.

2. Viz.ai LVO: AI-Powered Stroke Triage

Overview: The original Viz.ai **ContaCT** device received FDA De Novo classification on **13 February 2018** (DEN170073), as a Class II radiological triage and notification device ([FDA record](#)). It analyzes head CT angiography to notify specialists of suspected large-vessel occlusion. It operates alongside standard care and must not be relied on to make or confirm a diagnosis ([FDA decision summary and limitations](#)).

Evidence: The original authorization was a De Novo grant. The FDA review describes retrospective image-performance testing and notification-time evaluation. Special controls require verification and validation, standalone performance testing, effective-triage testing with appropriate datasets and uncertainty estimates,

software documentation, and labeling of intended users, population, compatible imaging, and limitations ([FDA review and special controls](#)). Later products or submissions should be assessed using their own authorization records.

Hassan and colleagues' 2022 study in *Stroke: Vascular and Interventional Neurology* used a single-center before-and-after design. Mean door-in to puncture times were 206.6 minutes before implementation and 119.9 minutes afterward, a difference of 86.7 minutes ($p < 0.001$), with higher adequate reperfusion rates ([original study](#)). This observational association does not establish causation or a mortality benefit.

The multicenter VALIDATE study measured arrival-to-neurointerventionalist notification, a separate endpoint. Its [2024 peer-reviewed report](#) should not be treated as confirmation of the single-center puncture-time result.

Regulatory and Clinical Lessons: Observational workflow evidence can inform evaluation, but confounding and changes in standard care require attention. **Hassan and colleagues**, rather than Weissman and colleagues, called for larger multicenter prospective studies to corroborate their findings ([study limitations](#)).

3. Mammography Screening (MASAI Trial)

Overview: Mass screening for breast cancer via mammography has historically been done by human radiologists. Recent European trials have begun to test AI as a secondary reader or screener. The MASAI trial in Sweden randomized over 100,000 women to either standard double-reading or AI-assisted reading.

Results: The 2026 MASAI analysis demonstrated **non-inferiority for interval cancer rates**, with a prespecified 20% margin. Rates were 1.55 versus 1.76 per 1,000 participants for AI-supported screening and standard double reading, respectively; the proportion ratio was **0.88 (95% CI 0.65–1.18; $p = 0.41$)**. The numerical 12% difference does not establish superiority for reducing interval cancers ([primary trial report](#); [trial investigators' university summary](#)). Interval cancers are cancers diagnosed between screening rounds, not a synonym for later-stage disease.

Stage and biological subtype were reported separately and **descriptively**: T2+ interval cancers numbered 38 versus 48, while non-luminal A interval cancers numbered 43 versus 59. These counts should not be combined into a single advanced/aggressive-cancer reduction or interpreted as proof of causation. Sensitivity was higher with AI support (80.5% versus 73.8%); specificity was 98.5% in both groups ([reported endpoints](#)). Mortality was not an endpoint of this analysis, so a survival benefit was not established.

Evidence Implications: The trial illustrates how a large prospective study can evaluate screening performance beyond a retrospective accuracy dataset. Implementation decisions should consider the prespecified endpoints, uncertainty, local workflow, and generalizability rather than assuming an unmeasured mortality benefit.

Challenges and Considerations

Throughout evidence generation, numerous challenges must be addressed:

- **Data Quality and Bias:** As noted, datasets often suffer from imbalance, limited diversity, and annotation error (www.sciencedirect.com). Many reviews report *spectrum bias* and overfitting in the literature (www.sciencedirect.com). To satisfy regulators, developers must curate representative cohorts and

transparently report dataset origins. Cross-ethnic validation is increasingly considered mandatory.

- **Explainability vs. Evidence:** Unlike a drug, an AI's internal logic may not be human-readable. Regulators still require justification of outputs. For example, the FDA's 2022 AI guidance suggests including clinical performance assessment and explaining algorithm decision thresholds ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Claiming a “black box” exempts evidence is not acceptable. Instead, developers should complement AI results with physician oversight or at least human review of a subset of cases for sanity-checking.
- **Performance Drift and Versioning:** AI models may degrade over time or with new data types (e.g., new scanner models, novel pathogens). Continuous performance monitoring is needed. FDA's [final PCCP guidance](#) provides a submission-based approach for planned modifications to AI-enabled device software functions. Post-market studies should track performance metrics over time.
- **Equity and Safety:** Emerging regulations (EU AI Act, FDA initiatives) demand demonstration of model fairness across subgroups. Studies should stratify results by age, sex, race to look for disparities. Documenting steps taken to detect and mitigate bias (as mandated by the AI Act (www.osborneclarke.com)) may become part of the evidence package.

Discussion and Future Directions

The landscape of evidence requirements for AI diagnostics is dynamic. Key future directions include:

- **Regulatory Evolution:** Under the [Commission's current timeline](#), the AI Act entered into force on 1 August 2024, generally applied from 2 August 2026, and applies high-risk Annex I regulated-product rules from 2 August 2028. Existing medical-device duties continue in the meantime.
- **Continuous Learning Systems:** Evidence for model changes should be planned throughout the device life cycle. FDA's [final PCCP guidance](#) was originally issued on 4 December 2024 and revised on 18 August 2025. FDA reviews the planned changes, validation and implementation methods, and impact assessment within the device's marketing submission. This framework is not a blanket authorization for unrestricted updates.
- **Converging Data Ecosystems:** Federated learning and big EHR datasets may allow more efficient validation. Future evidence generation might leverage large health networks (e.g. MIMIC, national registries) for multi-site validation without centralized datasets, addressing privacy and generalizability simultaneously.
- **Integration with Clinical Workflow:** Evidence generation will increasingly involve human factors. For AI to be accepted, studies must also show how it integrates seamlessly with clinician workflow. The technology is evolving toward more interpretability (e.g. heatmaps, explanations) to aid acceptance. Future guidelines may formalize the need for demonstration of UI/UX studies.
- **Economic and Ethical Outcomes:** Beyond clinical metrics, evidence on cost-effectiveness and ethical compliance will shape adoption. For instance, NICE and other HTA bodies expect health impact modeling. Transparency (open models or published algorithms) may become a factor in trustworthiness.

- **Public Trust and Liability:** Ultimately, robust evidence helps build trust among patients and deflects liability fears. Case law is emerging on AI misdiagnosis; evidence of rigorous validation may be a key defense. Regulators may even consider requiring public reporting of AI performance.

In conclusion, AI diagnostic tools hold intense promise but require commensurately rigorous clinical evidence for their claims. Regulators worldwide are moving towards standards that emphasize *representative data, performance transparency, and demonstrable patient benefit*. Early adopters like IDx-DR and Viz.ai demonstrate that well-designed studies—preferably prospective and with hard clinical endpoints—are feasible and valuable. As Weissman et al. assert, the **biggest barrier is evidence**: tools are being developed rapidly, but *concerns remain due to “investments outpacing evidence production”* ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/35811111/)). Closing this gap will demand collaboration: regulators must continue to clarify expectations, and developers must invest in high-quality trials and real-world studies. By prioritizing robust evidence—guided by emerging frameworks and the lessons of early AI diagnostics—the medical community can ensure that AI aids rather than undermines patient care.

Conclusion

Artificial intelligence diagnostic tools may support clinical expertise, but evidence must substantiate the intended use and claimed benefit. Developers should distinguish diagnostic accuracy from effects on outcomes and workflow, and assess what each regulatory pathway requires. FDA's device review evaluates study appropriateness for the device's intended use and technological characteristics ([FDA AI-enabled device resource](#)). EU medical-device obligations and the phased AI Act requirements should be assessed separately using the [current Commission timeline](#).

Our review of current literature and examples shows that well-designed evidence can clear the path for AI adoption. The IDx-DR and MASAI trials illustrate how prospective study data can meet regulatory standards and influence clinical practice ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/35811111/)) (ecancer.org). The Viz.ai experience shows that carefully collected real-world data can validate effectiveness in urgent-care settings ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/35811111/)). Yet we also highlight prevalent shortcomings: most published AI studies remain retrospective and unstandardized ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/35811111/)) (www.sciencedirect.com). To rectify this, we emphasize adherence to reporting guidelines (CONSORT-AI, DECIDE-AI, STARD-AI, etc.) and involvement of diverse, clinically realistic datasets. Stakeholders from academia, industry, and regulatory agencies must collaborate on iterative, transparent evaluation frameworks ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/35811111/)) (www.sciencedirect.com).

Looking forward, evidence generation for AI will have to keep pace with the technology's life cycle. Adaptive AI systems, advances in data sharing, and evolving patient safety concerns mean that evidence will be an ongoing commitment rather than a one-time hurdle. Incentives for post-market studies, cross-border data initiatives, and perhaps independent certification bodies could all play a role in ensuring continuous validation.

In sum, the clinical evidence requirements for AI diagnostic tools are stringent but fair: they aim to ensure that **only demonstrably safe, effective, and unbiased AI systems enter the clinic**. As AI becomes increasingly autonomous and widespread, these requirements will crucially balance innovation with patient protection. By

committing to rigorous, evidence-based validation today, developers and healthcare systems can help AI diagnostics achieve their potential to improve outcomes, reduce costs, and enhance the quality of care for all patients.

Tables

Region/Jurisdiction	Regulatory Body / Framework	Pre-market Evidence Requirements	Additional Notes
USA (FDA)	FDA (SaMD/MDR/MDUFA)	Risk-based pathways: PMA (PMA-level evidence), 510(k) (predicate equivalence, often limited clinical data) (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). Clinical studies expected where no predicate.	FDA provides AI/ML-specific guidance (e.g. GOOD ML practice, software precert), and mandates focused review of safety/effectiveness for cleared devices (www.fda.gov).
EU (MDR/IVDR)	MDR/IVDR conformity assessment; AI Act where applicable	Clinical or performance evaluation under the applicable medical-device regime. Qualifying regulated-product AI systems also face AI Act requirements under its phased timeline (AI Act).	Existing MDR/IVDR duties remain applicable; the current Commission timeline places Annex I high-risk AI rules at 2 August 2028 (timeline).
UK (MHRA/NICE)	Great Britain: UK MDR 2002/UKCA and conditional CE recognition; Northern Ireland: EU MDR/IVDR	Great Britain's UKCA requirements derive from earlier EU directives. CE recognition depends on device type, applicable legislation, and transitional conditions (MHRA guidance).	Northern Ireland requires CE marking, with UKNI where applicable. NICE ESF supports evaluation and purchasing decisions and is separate from regulatory approval (NICE framework).
China (NMPA)	NMPA (CFDA)	Rule-based classification: Many AI diagnostics considered Class III. Clinical trial often required for novel tools (pmc.ncbi.nlm.nih.gov). Emphasizes registration of algorithm details.	Chinese guidance is evolving; 59 AI devices approved by mid-2023 (pmc.ncbi.nlm.nih.gov). Focus on data security and ethics, but detailed clinical trial frameworks are less established globally.

Table 2. Overview of regulatory frameworks and evidence expectations for AI diagnostic devices. "Pre-market Evidence" summarizes general requirements per regulatory pathway. Sources: Weissman et al. (FDA insights) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)); EU MDR/AI Act guidance (www.osborneclarke.com) (www.osborneclarke.com); NICE framework ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)); Liu et al. on China ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

Figures: (Not included in text)

- (A) Flowchart of evidence generation stages for AI diagnostics (from in silico to RCT).
- (B) Example ROC curves and calibration plots from a hypothetical AI diagnostic trial.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://pubmed.ncbi.nlm.nih.gov/articles/PMC12339208/>
2. <https://www.osborneclarke.com/insights/eu-guidance-details-how-technical-and-clinical-standards-converge-ai-medtech>

3. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
4. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN180001>
5. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf
6. <https://pubmed.ncbi.nlm.nih.gov/41620232/>
7. <https://www.sciencedirect.com/science/article/pii/S2950489926000011>
8. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices?trk=article-ssr-frontend-pulse_little-text-block
9. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11410966/>
10. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8236783/>
11. <https://www.sciencedirect.com/science/article/pii/S2589750022000309>
12. <https://www.fda.gov/media/109618/download>
13. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
14. https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk?hss_channel=lcp-11318071
15. <https://www.nice.org.uk/what-nice-does/digital-health/evidence-standards-framework-esf-for-digital-health-technologies>
16. <https://www.lunduniversity.lu.se/article/ai-support-breast-cancer-screening-fewer-missed-cancer-cases>
17. <https://www.bmj.com/content/385/bmj-2023-078378>
18. <https://www.nature.com/articles/s41591-025-03953-8>
19. <https://pmc.ncbi.nlm.nih.gov/articles/PMC6550188/>
20. <https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/new-medical-services-and-new-technologies>
21. <https://www.fda.gov/media/166704/download>
22. <https://www.viz.ai/news/large-real-world-multi-center-study-demonstrates-viz-ai-platform-saves-critical-minutes-in-stroke-care>
23. <https://www.frontiersin.org/journals/stroke/articles/10.3389/fstro.2024.1381930/full>
24. <https://www.medtechdive.com/news/idx-publishes-ai-test-data-that-supports-landmark-approval/531188/>
25. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12864748/>
26. <https://www.cms.gov/files/document/12220-ops-final-rule-cms-1736-fc.pdf>
27. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN170073>
28. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN170073.pdf
29. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12778824/>
30. <https://ecancer.org/en/news/27721-ai-supported-mammography-screening-results-in-fewer-aggressive-and-advanced-breast-cancers-finds-full-results-from-first-randomised-controlled-trial>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.