AI SaMD Clinical Evidence: Real-World Performance Studies

By Adrien Laurent, CEO at IntuitionLabs • 4/14/2026 • 40 min read

samd | ai diagnostics | clinical evidence | real-world evidence | post-market surveillance | fda guidance | performance drift | medical software

Executive Summary

The maturation of **Software as a Medical Device (SaMD)** – particularly artificial intelligence (AI)-driven diagnostic tools – has introduced unprecedented opportunities and challenges for healthcare. Unlike traditional medical devices, AI-based SaMD can be continuously updated and may exhibit performance that shifts once deployed in diverse clinical settings. Consequently, robust *real-world evidence* (RWE) is crucial to ensure such tools are safe, effective, and equitable across patient populations. Historically, regulatory clearances have relied heavily on retrospective or simulated data. However, regulators and experts increasingly emphasize the need for prospective and post-market studies that measure actual clinical performance and capture potential *performance drift* (changes in accuracy over time or across settings) ([1] pmc.ncbi.nlm.nih.gov) ([2] www.fda.gov).

This report presents an in-depth analysis of clinical evidence generation for AI diagnostic SaMD, with a focus on **real-world performance studies**. We begin with historical and regulatory context: defining SaMD, summarizing global guidelines (FDA, IMDRF, EU MDR, G7, etc.), and outlining the unique features of AI tools (continuously learning, data dependence). We then explore study design strategies, from pre-market validation (retrospective studies, bench tests) to real-world clinical evaluations (prospective observational studies, pragmatic trials, registries, and continuous monitoring). Key methodological issues are examined, such as defining appropriate performance metrics (sensitivity, specificity, user-outcomes), ensuring representative data, and managing statistical challenges (sample size, bias, interoperability).

Several **case studies** illustrate these concepts in practice. For example, a recent *prospective* multi-phase implementation of an AI system for mammography screening in Hungary demonstrated that adding an AI reader yielded **0.7–1.6 extra cancers detected per 1,000 screenings** with only minimal increases in unnecessary recalls ([3] www.nature.com). In ophthalmology, the FDA-authorized IDx-DR system underwent a pivotal *prospective trial* in primary care settings, achieving **87.2% sensitivity and 90.7% specificity** for referable diabetic retinopathy ([4] www.nature.com). In contrast, a single-center *retrospective* study of an AI-assisted endoscopy system for gastric cancer found very high sensitivity (97.6%) but low specificity (45.8%) when used in practice ([5] www.mdpi.com). These examples highlight how real-world deployment can reveal different trade-offs (e.g. false-positives) than initial validation studies.

We also review the emerging regulatory and technical frameworks for AI SaMD **post-market surveillance**. Notably, the FDA's Digital Health Center of Excellence has solicited public input on measuring real-world AI performance ([2] www.fda.gov), reflecting a shift toward continuous quality monitoring. International guidelines (e.g. from a G7 expert panel) call for *life-cycle* driven evaluation: transparent reporting, ongoing monitoring, and risk management throughout a product's life (www.gov.uk) (www.gov.uk). Key challenges include detecting *data drift* and *concept drift*, addressing biases across subpopulations, and integrating clinician feedback. Technical solutions—such as federated learning, real-time data quality checks, and advanced causal inference methods—are under development but require further validation ([6] pmc.ncbi.nlm.nih.gov) ([7] pmc.ncbi.nlm.nih.gov).

Conclusions: AI-enabled diagnostic tools hold great promise for improving care, but only if they deliver reliable performance in real clinical settings. Our analysis underscores that pre-market testing alone is insufficient. Regulators, developers, and healthcare organizations must collaborate to implement robust real-world performance studies. This includes prospective clinical trials, continual performance monitoring (with planned metrics and triggers for action), and transparency about model updates. As summarized in international guidance, clinical evaluation of AI SaMD should be "rigorous, independent, continuous and proportionate" (www.gov.uk). Investing in such evidence generation will help ensure that AI diagnostics are safe, effective, and equitable, ultimately fulfilling their potential to improve patient outcomes.

Introduction

In recent years, artificial intelligence (AI)–driven software applications have begun to transform medical diagnostics and decision support. These AI tools, often implemented as **Software as a Medical Device (SaMD)**, use algorithms (especially machine learning models) to analyze clinical data – such as radiographic images, pathology slides, physiological signals, or electronic health records – and provide diagnostic or triage outputs without direct human interpretation of the raw data. By **SaMD**, we mean software that is clinically intended and performs a medical function independent of any specific hardware device ([8] pmc.ncbi.nlm.nih.gov) ([9] www.imdrf.org). The International Medical Device Regulators Forum (IMDRF) defines SaMD in this way to distinguish it from software that is integral to a physical medical device. Crucially, SaMD can be distributed on general-purpose computers, smartphones, or cloud platforms, making it highly scalable.

AI-based SaMD for **diagnostics** (sometimes called AI-assisted diagnostics) has proliferated particularly in imaging specialties. For example, computer-aided detection (CADe) and computer-aided diagnosis (CADx) algorithms for radiology and pathology have long been under development. Classic examples include mammography CADe systems first cleared in the late 1990s. By 2016, fully 90% of U.S. radiology centers were using CAD for mammogram screening ([10] pmc.ncbi.nlm.nih.gov). However, later *real-world* analyses revealed that CAD did not actually improve cancer detection in practice; it increased sensitivity at the cost of far more false positives, leading to substantial unnecessary workups ([11] pmc.ncbi.nlm.nih.gov). This sobering lesson has underscored the need to thoroughly validate new AI tools in the actual clinical contexts where they will be used. Today, advances in deep learning have enabled new algorithms in fields as diverse as dermatologic imaging, ophthalmology (retinal scans), cardiology (ECGs, echocardiograms), gastroenterology (endoscopic images), and beyond ([12] www.mdpi.com) ([13] www.nature.com). Some systems are now deployed globally, but many questions remain about how to best evaluate their true benefit and risk.

The **unique nature of AI SaMD** poses several challenges for evidence generation. Unlike a fixed hardware device or traditional drug, AI software may be updated frequently (even adaptively) as more data become available.It typically relies on large training datasets and complex algorithms, which may exhibit unpredictable behavior when data distributions shift. In clinical use, changes in patient populations, equipment, or user behavior can all impact an AI model's performance. Regulatory bodies have thus begun to emphasize a **total product lifecycle** approach: rigorous pre-market evaluation is necessary but not sufficient, and real-world post-market monitoring is also crucial (www.gov.uk) ([2] www.fda.gov). In other words, simply clearing an AI tool based on retrospective studies is not a guarantee of ongoing safety or effectiveness once the tool faces the complexities of routine care.

This report delves deeply into **clinical evidence generation for AI SaMD**, with an emphasis on designing and conducting **real-world performance studies** for AI diagnostic tools. We first outline the current regulatory landscape and historical context of SaMD. We then examine the design of clinical studies – how to validate an AI device pre-market and how to assess it after deployment. We review data sources (electronic health records, registries, image databases) and statistical considerations (bias, sample size, metrics). We highlight illustrative case studies of AI diagnostic tools in different domains, analyzing how these studies were structured and what they revealed. Finally, we discuss emerging themes: managing performance drift, ensuring fairness, and evolving regulations. Throughout, we draw on the latest guidelines, peer-reviewed studies, and expert commentaries to support our analysis ([2] www.fda.gov) (www.gov.uk) ([3] www.nature.com) ([4] www.nature.com).

Background and Regulatory Context

Evolution of SaMD and AI Devices

The concept of SaMD has gained prominence as digital health innovations accelerated. In 2013, the IMDRF formed a working group on SaMD to harmonize international guidance. By 2017–2018, major regulatory agencies had published frameworks specific to AI-driven medical software. For example, in April 2019 the U.S. Food and Drug Administration (FDA) issued a series of discussion papers and draft guidance documents outlining its proposed approach for **AI/ML-**

driven SaMD ([8] pmc.ncbi.nlm.nih.gov). Key themes included managing algorithm modifications and ensuring controls over training data and outputs. The IMDRF's SaMD working group completed its formal guidance on clinical evaluation in 2017 (IMDRF/SAMD WG/N41) which stressed evidence generation but was not AI-specific. Alongside, the 21st Century Cures Act (2016) in the U.S. attempted to streamline digital health approvals, though subsequent clarification was needed for SaMD.

Other jurisdictions followed suit. The EU's Medical Device Regulation (MDR 2017/745, effective 2021) explicitly encompasses software and often classifies medical software as Class IIa or higher, requiring comprehensive clinical evaluation and post-market surveillance. The prospective EU **AI Act** (agreed in 2023, phased in by 2025 onwards) goes further by categorizing medical AI as "high-risk," mandating robust testing, transparency, and lifecycle monitoring. The UK's Medicines and Healthcare products Regulatory Agency (MHRA) has issued guidance on AI in medical devices, aligned with G7 principles of safety, effectiveness and ethical oversight (www.gov.uk). In Asia, Japan's PMDA and China's NMPA have introduced requirements for AI software validation, often mirroring FDA/IMDRF principles.

Across agencies, the consensus is growing: *validating AI SaMD demands a lifecycle approach*. As regulators note, AI models "in deployment become more frequently updated – and, in due course, 'continuously-learning' – assurance of performance will require continuous monitoring" (www.gov.uk). This reflects lessons learned from early CAD systems: approval did not guarantee real-world benefit ([11] pmc.ncbi.nlm.nih.gov). The US example of mammography CAD illustrates the need for lifecycle thinking. Despite achieving FDA clearance on retrospective tests, widespread use of mammography CAD ultimately showed **no improvement** in cancer diagnosis outcomes, and even led to excess false positives and costs ([11] pmc.ncbi.nlm.nih.gov). The Centers for Medicare & Medicaid Services (CMS) eventually withdrew extra reimbursements for CAD, acknowledging it had provided no real benefit ([14] pmc.ncbi.nlm.nih.gov).

Regulatory Guidance on Evidence

Major regulatory bodies have begun to articulate expectations for clinical evidence. In the U.S., FDA's guidance on SaMD (2017) and on modifications to AI/ML devices (proposed 2019) emphasize transparency of training data and continuous learning protocols. In January 2025, the FDA released a draft guidance on **AI/ML SaMD Lifecycle Management**, recommending that manufacturers provide documentation of design, training, and risk mitigation throughout the product's life ([15] www.fda.gov). The agency is particularly interested in how developers will measure real-world performance ("postmarket performance" in draft guidance) ([15] www.fda.gov).

The FDA has famously created an **AI/ML Medical Device Action Plan** that includes piloting new pathways for adaptive AI and advanced testing methods. Crucially, in late 2022 the FDA announced a **Request for Public Comment** (RFI) on *"Measuring and Evaluating AI-Enabled Medical Device Performance in the Real World"* ([2] www.fda.gov). This RFI explicitly recognized that "many AI-enabled medical devices are evaluated primarily through retrospective testing or static benchmarks," which are *not* sufficient to predict how devices behave in dynamic clinical use ([2] www.fda.gov). It asked stakeholders to propose metrics, monitoring processes, and infrastructure for real-world performance evaluation. This shift signals regulator interest in moving beyond initial approval to ongoing assurance of efficacy and safety.

International guidance similarly highlights continuous monitoring. For example, a 2021 G7 expert panel report laid out **key principles** for AI/ML device evaluation, stressing that assessment should be "rigorous, independent, continuous and proportionate" and include post-market surveillance (www.gov.uk) (www.gov.uk). It recommends that clinical trials or studies "balance collecting robust evidence with the need to minimise risk to patients and workflow disruption," and explicitly supports retrospective "shadow mode" testing when appropriate (www.gov.uk). The panel also urges transparent reporting of diagnostic studies using standards like **CONSORT-AI** and **STARD-AI** for trials and diagnostic accuracy studies, respectively (www.gov.uk). In sum, authoritative guidelines now expect a cycle of evidence generation that extends from lab validation to real-world clinical study to ongoing audit.

Current Landscape of AI SaMD

The rapid growth of AI tools for diagnostics is often illustrated by numbers. For instance, in radiology alone over 100 AI algorithms had received FDA clearance by 2021 ([16] www.sciencedirect.com), and this number is rising sharply. A 2024 analysis catalogued **309 AI/ML SaMD approvals by the FDA** (2014–2023) across specialties ([17] pmc.ncbi.nlm.nih.gov). These were categorized by intended use: about 42% were computer-aided detection (CADe) algorithms (flagging possible abnormalities), 6.8% were diagnostic (CADx), 21.7% triage tools, and the rest for workflow optimization or image acquisition ([17] pmc.ncbi.nlm.nih.gov). This proliferation underscores the urgency for rigorous evidence: with dozens of new tools entering practice each year, hospitals and clinicians must know which ones truly improve care.

Many of these cleared AI tools relied largely on **retrospective dataset evaluations** for evidence. Published analyses indicate that most early FDA-submitted AI applications were supported by historical image reviews or small validation cohorts ([18] pmc.ncbi.nlm.nih.gov). In some cases, the details of study sites or sample sizes were not fully disclosed in summary documents, limiting transparency. Experts have argued that such pre-market testing, while a necessary first step, does not guarantee generalizability. For example, a model trained and validated on data from a specialized academic hospital may not perform the same way in community clinics or among different ethnic groups. Consequently, leading voices now insist on **broad, prospective validation** across diverse clinical settings.

In parallel, the field is also recognizing the limitations of static validation. The “FDA-cleared” label no longer implies a one-time guarantee. AI models can degrade due to **data drift** (changes in patient populations or inputs) or **model drift** (updates in clinical practice) ([19] www.fda.gov). The FDA’s RFI defines data drift broadly (changes in inputs/outputs leading to bias or reduced reliability) and cautions that ongoing monitoring is needed to detect such drifts ([19] www.fda.gov). Similarly, a 2025 report in *NEJM AI* highlighted how clinical interventions can confound evaluation: if an AI alert leads to treatment that prevents the predicted outcome, naïve monitoring may wrongly infer the model is over-predicting risk ([6] pmc.ncbi.nlm.nih.gov). These subtleties reinforce that **post-market surveillance of AI devices** will require novel strategies and careful interpretation.

In summary, as AI diagnostics become more prevalent, the demand for **real-world performance evidence** has intensified. Regulatory expectations are evolving accordingly. The rest of this report examines how to generate this evidence: what study designs and data sources are suitable, what lessons can be drawn from recent cases, and how stakeholders can work together to ensure these promising technologies deliver real clinical value.

Clinical Evidence Generation for SaMD

To evaluate an AI diagnostic SaMD, stakeholders seek evidence on two key aspects: **technical performance** (does the software correctly analyze data, e.g. accurately detecting pathology on a match dataset?) and **clinical performance/effectiveness** (does its use lead to improved patient outcomes or decision-making? Does it work safely in practice?). Both must ultimately be assessed in the context of intended use (population, setting, user). Importantly, the approach differs somewhat from clinical trials for drugs or hardware: software can be tested in silico or in simulation, updated post-release, and its interaction with users is dynamic. Below we outline the spectrum of evidence generation.

Pre-market Evaluation

Retrospective Validation Studies

Most AI SaMD begin with **retrospective performance studies**. These typically involve applying the algorithm to existing data (e.g. de-identified imaging archives or EHRs) that have been labeled by human experts (the “ground truth”). Performance metrics (sensitivity, specificity, AUC, etc.) are calculated by comparing the AI outputs to these labels. For

instance, a chest X-ray CAD algorithm might be tested on a curated dataset of 10,000 annotated images collected from multiple hospitals. Such retrospective testing is relatively quick and inexpensive, and allows initial validation without patient risk. Many FDA submissions use this approach to demonstrate non-inferiority or superiority to standard practice ([18] pmc.ncbi.nlm.nih.gov) ([16] www.sciencedirect.com).

Retrospective studies are valuable for technical validation and ensuring the AI algorithm meets basic accuracy requirements. They enable large sample sizes (sometimes >100,000 cases) if multi-center datasets are available ([20] www.nature.com) ([16] www.sciencedirect.com). The G7 principles note that retrospective studies may *explore the possibility* of evaluation "to minimise disruption" (by, e.g., running the AI in shadow mode) (www.gov.uk). However, retrospective designs have limitations: they may not mimic how the tool is used in the clinic. For example, images selected retrospectively might be of higher quality than routine scans, or the labeler consensus process might not reflect real-time decision contexts. There is also the risk of *spectrum bias*: if the test set is not representative, the calculated performance will not generalize. For AI, performance on retrospective data generally needs to be supplemented by more rigorous prospective or comparative evaluation before one can be confident about its clinical impact ([21] www.nature.com) (www.gov.uk).

Prospective Clinical Studies

To address real-world performance, **prospective clinical trials or studies** are crucial. In these, the AI tool is deployed as part of the clinical workflow (either live or in parallel) and outcomes are measured prospectively. Study designs vary:

- **Randomized controlled trials (RCTs):** The gold standard, rarely used for SaMD due to cost and logistics, but some do exist. For example, an RCT could randomize patients (or clinics) to use the AI tool versus usual care, then compare outcomes (e.g. cancer detection rates, decision accuracy, patient outcomes).
- **Non-randomized prospective cohort studies:** More common. Patients or cases are consecutively enrolled and the AI tool's recommendations are observed. Outcomes may be compared to historical controls or a concurrent control group, or measured against ground truth established by follow-up.
- **Switched or before-after designs:** Some studies compare performance metrics before and after AI deployment in the same clinics (though changes over time may confound).
- **Enhancement of existing workflows:** A hybrid approach is to run the AI alongside standard practice (shadow mode) for a period, then evaluate how often it identified cases missed by clinicians. The G7 principles explicitly mention "silent mode" or "shadow mode" evaluations (www.gov.uk) as lower-risk ways to test a new tool before full integration.

Prospective studies directly measure how the AI performs in the intended environment. The 2023 *Nature Medicine* breast screening study is a prime example: at multiple screening centers, AI was introduced as an additional reader in real-time, and metrics like cancer detection and recall rates were tracked across phases ([3] www.nature.com). Similarly in diabetic retinopathy (IDx-DR), the pivotal trial was prospective – patients in primary care were screened with AI, and then confirmed by ophthalmic exams ([4] www.nature.com). These prospective designs revealed insights unavailable from retrospective data alone (see Case Studies below).

However, prospective trials are complex. They require workflow integration, training of staff, clear definitions of endpoints, and often ethics oversight. Enrolling large numbers of patients can be slow. Regulators acknowledge this trade-off: the G7 report suggests balancing robust evidence against patient risk and workflow disruption (www.gov.uk). In practice, a staged approach is common: first retrospective validation, then limited prospective pilots, then larger rollouts if performance is promising.

Clinical Endpoints

Depending on the tool's purpose, endpoints in these studies vary. For a diagnostic support algorithm, classical outcomes include technical accuracy (sensitivity, specificity, AUC for detecting disease) and diagnostic yield (e.g. cancers detected

per screened population). Equity and subgroup analyses (performance by age, sex, race/ethnicity) are increasingly required. Some studies also track *downstream* clinical outcomes: for example, does using the AI change time to treatment, treatment decisions, or patient morbidity/mortality? The G7 guidance emphasizes inclusion of patient-centered outcomes (e.g. how the tool affects patient experience or quality of life) where possible (www.gov.uk). User-centric metrics are also important: how does the AI affect clinician workflow, decision confidence, or system efficiency? Real-world studies should ideally record false positives and negatives, user overrides, and follow-up procedures triggered by the AI.

Any evaluation should pre-define key performance indicators. For example, in an immuno-cytology screening AI, one might measure *per-patient* sensitivity (at least one disease flagged) and *recall rate*. In image analysis, metrics might be per-image or per-lesion. Standardized reporting guidelines (CONSORT-AI for trials, STARD-AI for diagnostic accuracy, TRIPOD-AI for predictive models) are recommended for clarity and comparability (www.gov.uk). Using these standards ensures transparency: for instance, listing how patients were selected, how data were partitioned, and how endpoints were adjudicated.

Post-market and Real-World Studies

Upon regulatory clearance, AI SaMD is often used clinically, enabling the collection of real-world data. Post-market studies can be either planned at launch (as a condition of approval) or conducted voluntarily by developers to demonstrate value or detect issues. These studies exploit data generated during actual use, with minimal interference.

Continuous Monitoring and Post-market Surveillance

Regulators expect a **Total Product Lifecycle (TPLC)** approach: the manufacturer is responsible for ongoing monitoring and reporting of safety/effectiveness throughout deployment (www.gov.uk). This includes routine tracking of model performance (via defined metrics) and adverse event reporting if the AI contributes to errors. For example, an AI radiology tool vendor might monitor its false positive rate over time as more sites adopt it. The FDA draft guidance calls for documentation of post-market performance checks and predefined methods to assess drift ([2] www.fda.gov) ([15] www.fda.gov).

Some companies and health systems have begun creating monitoring infrastructures. These may include data pipelines comparing AI outputs to ground truth as cases follow up (for instance, checking whether nodules flagged as benign by AI later turned malignant). Tools like automated data audits can flag shifts in input data characteristics (e.g. a new imaging protocol). The challenge is choosing triggers for action: when should a small drop in sensitivity prompt a model update or retraining? The FDA's RFI explicitly asks about "monitoring triggers and response protocols" (Questions 4a–4b) for identifying when performance degradation warrants intervention ([22] www.fda.gov).

Real-World Evidence (RWE) Studies

Beyond routine monitoring, some post-market evidence is gathered through formal RWE studies. These often use electronic health records (EHR), claims databases, or registries to evaluate AI performance retrospectively. For example, a hospital network might retrospectively analyze one year of consecutive imaging with AI assistance, measuring outcomes like diagnostic accuracy or missed cases. RWE can span larger populations than single-site trials, improving generalizability. It also allows flexibility: if an AI tool gains traction in one region, data from multiple sites can be pooled to assess consistency.

A key concept is the **observational performance study**. These can be prospective registries (clinicians record data in parallel with AI use), or retrospective chart reviews across sites. For instance, a multi-center registry might capture how an AI-based endoscope was used in routine colonoscopies, tracking adenoma detection rates with AI versus historical controls. The npj Digital editorial on real-world AI performance calls for "longitudinal studies of AI in practice over time" as

a critical research need ([23] www.nature.com), emphasizing that RWE should address scalability and generalizability across institutions and care settings.

Real-world studies face challenges such as data quality and confounding. EHR data can be incomplete or biased if clinicians change behavior. The NEJM AI report by Ansari et al. points out a subtle problem: when an AI model successfully prompts an intervention (e.g. early treatment), the subsequent outcome may improve, making the model *appear* to under-perform if one only looks at raw outcomes ([6] pmc.ncbi.nlm.nih.gov). This "lab versus life" paradox means that RWE analysts must consider the feedback loop between AI recommendations and clinical actions.

Post-market Modification Studies

AI SaMD often undergoes updates after approval (for new data or improved algorithms). Regulatory policy for these modifications is evolving. Currently, minor updates (e.g., fixing software bugs or small performance tweaks) may only require **standalone software testing** by the vendor and notification to regulators, whereas major updates (changing intended use or algorithms substantially) might trigger another formal review or a new 510(k). A recent survey of FDA-cleared AI/CAD tools noted that post-market updates were almost always evaluated retrospectively ([24] pmc.ncbi.nlm.nih.gov). If the intended use didn't change, companies tended to use automated test methods; only when use cases changed did they do a reader study with human experts ([24] pmc.ncbi.nlm.nih.gov) ([25] pmc.ncbi.nlm.nih.gov).

This has led to calls for clearer guidance: the same survey suggests that industry and regulators should agree on when standalone testing is sufficient versus when new prospective studies are needed ([26] pmc.ncbi.nlm.nih.gov). The risk is that without standards, minor alterations could accumulate, drifting the model away from its originally validated state. Patients and clinicians need transparency: some authorities now recommend that major algorithm updates should be communicated to users and possibly undergo fresh evidence generation, especially if the updates affect performance in subgroups (www.gov.uk) (www.gov.uk).

Study Design Considerations

When planning real-world performance studies for AI SaMD, the following methodological factors must be addressed:

- **Population Representativeness:** The study cohort should reflect the intended user population. Does the study include the full spectrum of cases? For instance, if an AI tool is meant for community clinic use, validating it only in a tertiary hospital sample risks optimism bias. Stratify results by relevant subgroups (age, ethnicity, disease prevalence).
- **Comparator Definition:** What is the gold standard or reference? For diagnostic tools, this might be biopsy results, consensus panel determinations, or long-term follow-up. For predictive models, one must define outcomes precisely. The design should consider whether to compare AI vs expert, AI+expert vs expert alone, or AI vs historical rates.
- **Study Setting and Blinding:** Ideally, those assessing outcomes should be blinded to the AI result to avoid bias. In real-world deployment, this may not always be feasible, but prospective studies often use an independent adjudication committee.
- **Sample Size and Power:** Statistical planning is as crucial as for drug trials. If superiority or non-inferiority hypotheses are tested, sample size calculations must account for expected prevalence of conditions. Rare conditions may demand very large datasets. Sparse data in subgroups (e.g. pediatric patients) must be noted.
- **Workflow Impact:** Studies should consider how integrating AI affects workflow. For example, if reading a chest X-ray with AI takes longer due to interface differences, this may affect throughput. Mixed-methods studies (combining quantitative outcomes with user surveys) can capture usability issues.
- **Regulatory Endpoints:** For some SaMD, regulatory approval may hinge on certain benchmarks (e.g. IDx-DR had pre-specified targets of >85% sensitivity and >82.5% specificity ([4] www.nature.com)). When replicating studies in the real world, one should aim to meet or define the context for those benchmarks.

Collectively, these design elements ensure that real-world studies provide **actionable evidence**: not just "Is the AI model technically accurate?" but also "Does its use change practice or patient results for the better, without undue harm or

burden?”

Data Sources and Infrastructure

Real-world performance studies rely on diverse data types and often on innovative collection methods. Key sources include:

- **Electronic Health Records (EHRs):** Structured data (lab results, demographics, diagnoses, procedures) and unstructured data (clinician notes, imaging) can be mined. For example, an AI sepsis prediction tool might be monitored by examining EHR for reported sepsis cases, care interventions, and outcomes. Data linkage (e.g. matching imaging records to pathology results) is often needed.
- **Registries & Networks:** Disease-specific or procedure-based registries (e.g. stroke registries, cancer registries) offer curated datasets. If an AI tool is used clinically, it may be added as a field in a registry form. For instance, breast screening programs maintain registries of mammograms and outcomes, which can be used to retrospectively or prospectively assess AI tools as in the Hungarian example ([27] www.nature.com) ([3] www.nature.com).
- **Device & Software Logs:** AI systems often generate logs of predictions and inputs. These logs can be anonymized and analyzed to find trends in usage and errors. For example, an AI radiology tool might record the confidence scores and decisions for each exam, enabling aggregate performance monitoring.
- **Patient-Reported Outcomes:** Some AI tools aim to improve patient experience (e.g. AI-driven tele-dermoscopy might reduce anxiety). Surveys or digital questionnaires can capture outcomes like satisfaction or perceived quality of care.
- **Claims and Billing Data:** Though coarse, billing codes can identify diagnoses, procedures, or errors post hoc. If an AI reduces misdiagnoses, claims data might show fewer procedure codes for additional follow-ups.

Infrastructure is a practical consideration. Implementing real-world studies often requires partnerships with clinical sites to integrate data collection with routine care. Automated pipelines (possibly cloud-based) to de-identify and aggregate data are beneficial. Tools for *data quality assessment* are critical: e.g. the Emerging Health AI community has developed platforms like the “Metric Hub” to check dataset representativeness before an AI is built ([23] www.nature.com). When collecting EHR data across institutions, standards like FHIR (Fast Healthcare Interoperability Resources) or OMOP common data model help ensure comparability.

Data privacy and governance are paramount. Many real-world datasets contain personal health information. Consent (or proper legal frameworks) is needed for using patient data in studies. The industry is exploring federated analysis models to respect privacy: under federated learning or federated evaluation, each site analyzes local data and only shares aggregated metrics or model updates. This approach has been highlighted as a way to expand real-world testing while minimizing central data pooling risks.

Case Studies of Real-World Performance Studies

To illustrate how real-world evidence can be generated for AI diagnostics, we examine a few representative case studies across specialties.

Breast Cancer Screening (Radiology): Prospective AI Reader Trial

Context: A deep-learning system (Mia by Kheiron Medical) designed to flag potential cancers on screening mammograms was evaluated in Hungary ([3] www.nature.com). Historically, AI in radiology—like earlier CAD—had failed to improve outcomes in practice ([11] pmc.ncbi.nlm.nih.gov), so this prospective study aimed to rigorously test the new AI in live screening.

Study Design: The study was implemented in three phases at four screening sites (urban and rural). Initially, a single radiologist reviewed exams flagged "no cancer" by standard double-reading but flagged "possible cancer" by AI (a pilot phase). Then a multicenter pilot extended this with three radiologists/arbitrators reviewing AI-flagged discordant cases. Finally, a full live rollout incorporated AI flags into the final read by additional arbiters. This "additional-reader" workflow had AI act as a third reader in a primarily double-read program, alerting arbitrators to re-examine cases otherwise deemed normal ([28] www.nature.com) ([29] www.nature.com).

Population: Over the study period, tens of thousands of women (ages ~58) were screened routinely. The first pilot included 3,746 screens, the expanded pilot 9,112, and live rollout continued beyond these phases ([30] www.nature.com) ([31] www.nature.com).

Metrics and Outcomes: Key metrics were additional cancers detected and recall rates. The AI-assisted strategy detected **0.7–1.6 more cancers per 1,000 women** than standard double reading alone ([3] www.nature.com). Importantly, most of the additional cancers were invasive (83.3%) and small (≤10 mm diameter, 47%) ([32] www.nature.com), meaning they were early-stage and prognostically significant. The AI prompted only a small increase in recalls: 0.16–0.30% more women were called back for further testing ([3] www.nature.com). The positive predictive value (PPV) of recalls actually improved slightly (0.1–1.9% increase) despite the extra reads, suggesting the AI flags had high yield ([3] www.nature.com). In practical terms, the AI reviewed just 4–6% additional volume (flagged cases that humans might have missed) while achieving a net gain in early detections.

Insights: This real-world, prospective evidence showed that AI could be beneficial as a "safety net" reader in screening, aligning with its intended use. It demonstrated the feasibility of integrating AI into workflow with modest extra workload (Additional Table **1**). It also highlights the importance of prospective design: retrospective test misses do not directly translate to clinical impact without factors like arbitration processes. This study influenced practice, with one of the sites adopting the AI-assisted workflow as part of clinical routine. Had decisions been based only on retrospective validation, the nuanced trade-off of extra reads vs detected cancers might have been overlooked.

Table 1. Summary of AI-Assisted Mammography Study (Ng et al., Nature Med 2023)
Setting: 4 breast cancer screening sites (Hungary); double-reading workflow with AI as extra reader.
Phases: (1) Single-reader pilot, (2) multi-reader pilot, (3) live rollout with 3 independent arbitration readers.
Cases: ~3,700 screens (phase 1) + 9,100 (phase 2). Ongoing in phase 3.
Key Results: +0.7–1.6 additional cancers per 1,000 screens (versus double-reading alone); 83.3% of these were invasive, 47% ≤10mm.
Recalls: +0.16–0.30% absolute recalls; +0–0.23% unnecessary recalls.
PPV Change: Increased by 0.1–1.9% (higher PPV implies improved efficiency despite more reads).
Workload: 4–6% more readings needed overall (due to 7–11% of cases flagged by AI requiring extra arbitration).
Citation: Ng et al. (Nature Med 2023) ([3] www.nature.com) ([27] www.nature.com)

Diabetic Retinopathy Screening (Ophthalmology): Pivotal Prospective Trial

Context: IDx-DR was the first autonomous AI diagnostic system cleared by FDA (2018) for diabetic retinopathy (DR) screening ([4] www.nature.com). Unlike most AI tools, IDx-DR was intended to operate without specialist oversight: patients could get retinal photos in primary care, and the AI would autonomously decide if they should be referred to eye specialists.

Study Design: A multicenter prospective trial (NCT02963441) enrolled 900 patients with diabetes at primary care clinics. Each participant underwent retinal imaging by certified photographers, followed by imaging at the Wisconsin Fundus Photograph Reading Center as gold standard. The AI's output (mtmDR: more-than-mild diabetic retinopathy) was compared to the human grading of ETDRS (Early Treatment Diabetic Retinopathy Study) levels and macular edema presence ([4] www.nature.com).

Outcomes: The AI system achieved **87.2% sensitivity** and **90.7% specificity** for detecting >mild DR (pre-specified targets were >85% and >82.5%, respectively) ([4] www.nature.com). Its imageability rate (ability to get a gradable result) was 96.1%. These figures exceeded the FDA's threshold, enabling clearance. Notably, this high level of accuracy was achieved across a broad patient spectrum: median age 59, balanced gender, with 28.6% African American, 16.1% Hispanic ([4] www.nature.com).

Insights: The IDx-DR trial demonstrates a carefully controlled prospective evaluation in the intended use population (primary care diabetic patients). It represents a benchmark for evidence: a large, prospective, comparative study with rigorous endpoints that directly supported regulatory approval ([4] www.nature.com). For real-world impact, the fact that such an AI can match ophthalmologist grading in this context suggests it can expand access to screening. Subsequent real-world observational studies would be needed to see if screening rates or patient outcomes improve when using IDx-DR. This trial also highlights the challenge of maintaining standards: even <mild DR can have subtle signs, so rigorous image quality control (an element of the device) was essential.

Endoscopy for Gastric Cancer (Gastroenterology): Retrospective Single-Center Study

Context: AI-augmented endoscopy promises to catch early gastrointestinal cancers. Osawa et al. (2024) evaluated a commercial AI system for detecting early gastric cancer in a real clinical setting ([33] www.mdpi.com) ([34] www.mdpi.com).

Study Design: A single-center retrospective study included 47 patients (89 lesions) who underwent white-light endoscopy between March 2024–March 2025. Each lesion was repeatedly assessed by the AI during the endoscopy. A lesion was considered AI-positive if ≥50% of AI assessments said "consider biopsy". The investigators calculated sensitivity, specificity, PPV, etc., of AI vs the pathology of biopsied lesions ([35] www.mdpi.com). Logistic regression identified factors linked to false-positive AI detections.

Results: The AI achieved an extremely high **sensitivity of 97.6%** (catching nearly all early cancers) and a **negative predictive value of 95.7%**, but **specificity was only 45.8%** ([34] www.mdpi.com). In other words, it flagged many lesions as suspicious that were not cancer (false positives were common). The analysis further showed that certain conditions – like lesion size ≥30 mm, scars, or erosions – predicted false positives ([36] www.mdpi.com). Interestingly, the AI's output confidence stratified outcomes: lesions with high AI confidence had higher pathology-positive rates ([34] www.mdpi.com).

Implications: Unlike the previous examples, this was a retrospective real-world analysis, but in *actual clinical practice*. It underscores that AI tools may behave differently in messy reality than in initial development. The tool's design target (84.7% sensitivity, 58.2% specificity on retrospective video) had already highlighted a trade-off ([37] www.mdpi.com). In practice, sensitivity was even higher (possibly due to using the tool repeatedly on each lesion), but specificity was lower (thus many unnecessary biopsies). The study's suggestion – using confidence thresholds to modulate action – is a novel insight for implementation. It shows that real-world studies can guide how clinicians should interpret AI scores (not just binary outputs), to optimize benefit vs burden.

Table 2. Case Studies of AI Diagnostic Tools and Their Evidence Studies
Application & Tool
Mammography screening (AI as extra reader)
Ophthalmology: Diabetic Retinopathy (IDx-DR)

Table 2. Case Studies of AI Diagnostic Tools and Their Evidence Studies
GI Endoscopy: Gastric CA detection (gastroAI)
Previous CAD History: Mammography (no AI)

Table 2 summarizes these examples. Collectively, they illustrate a spectrum of evidence: from high-volume prospective deployments to smaller retrospective reviews. A striking contrast is seen between the mammography and endoscopy studies: the former added meaningful early detections with controlled minimal downside ([3] www.nature.com), whereas the latter, despite high sensitivity, risked many false alarms in practice ([34] www.mdpi.com). Meanwhile, the IDx-DR trial exemplifies rigorous prospective validation leading to broad regulatory acceptance ([4] www.nature.com). These cases highlight that **context matters** – patient population, workflow, and follow-up pathways all influence how an AI tool will perform for real patients.

Statistical and Methodological Considerations

Rigorous data analysis underpins credible performance studies. Key issues include:

- **Performance Metrics:** Traditional classification metrics (sensitivity, specificity, AUC) remain central. However, in practice negative and positive predictive values (NPV/PPV) often matter more clinically. For rare conditions, PPV may be low even with high sensitivity. When deploying AI as an assistive tool, one may also track metrics like *false positive rate per patient* or per screen. Equally important are **safety metrics**: e.g. rate of missed cases leading to adverse outcomes. The FDA RFI (Q1) asks what metrics should measure safety, effectiveness, reliability ([38] www.fda.gov). Some answers propose composite endpoints, such as *net benefit to patient* or thresholds of acceptable alert rates.

- **Statistical Design:** Hypothesis setting (superiority, non-inferiority) dictates sample size. SaMD trials may require multi-reader design, where multiple clinicians (with/without AI) interpret cases, increasing power. In reader studies, controlling for inter-reader variability is crucial (e.g. crossing over readers with/without AI). When using historical controls or “before-after”, statistical adjustments (propensity scores or regression) may help, but careful baseline matching is needed to avoid confounding. For example, if an AI was introduced in 2022, one should ensure 2021 data is comparable (same seasonality, similar patient mix).

- **Generalizability and External Validation:** Even if a study shows good performance, one must ask: will it hold elsewhere? A tool validated in one health system may falter in another with different equipment or disease prevalence. Best practice is to test on external datasets. The G7 document underscores that regulators should demand validation on independent datasets **“reflecting the diversity of the intended population and setting”** (www.gov.uk). This aligns with IMDRF guidance that SaMD evaluations should consider all factors (patient, site, data) that might influence performance.

- **Handling Imbalance:** Many medical conditions are low-prevalence. Datasets often have far more negatives than positives. This requires careful handling (e.g. stratified sampling, enrichment of cases, or use of appropriate statistical measures like precision-recall curves). When designing real-world studies, oversampling rare cases (or pooling multi-center data) may be necessary.

- **Subgroup Analysis:** Ethical and regulatory considerations demand that performance across demographic groups is examined to spot biases. This means studies must be powered to evaluate minorities or special subpopulations. For instance, an AI tool might work well overall but poorly in underrepresented racial groups if not properly trained. The UK-G7 guidance explicitly asks manufacturers to report model performance “for appropriate subgroups at the intersections of demographics and clinical status” (www.gov.uk).

- **Data Quality:** Garbage in, garbage out. Real-world data are often messier than clinical trial data. Missing values, variable imaging protocols, and transcription errors can degrade model performance. Studies should include methods to assess and mitigate data quality issues. For example, one may compare AI outputs on data from different imaging devices to check consistency.

- **Ethics and Privacy in Study Design:** Especially when trialing AI that may affect patient care, ethical oversight is needed. Some prospective studies use “silent mode” (clinicians do not see AI output) to avoid jeopardizing care. Patient consent for using their data (even de-identified) must be addressed. The G7 report points out the need to consider patient welfare and privacy throughout evaluation (“promote inclusiveness, fairness and transparency”).

In sum, designing real-world performance studies requires blending traditional biostatistics with novel approaches suited to software. Transparency in reporting (using CONSORT-AI, etc.) and preregistering study protocols (e.g. on clinicaltrials.gov) are also emerging best practices to avoid selective reporting.

Emerging Tools and Technologies for Performance Monitoring

Several technical strategies are under development to support AI SaMD evaluation and monitoring:

- **Continuous Integration/Deployment Pipelines:** Some AI developers are creating DevOps frameworks for health AI where model performance metrics are continuously tracked. For example, A/B testing within a hospital, or canary deployments where updates are rolled out to a subset of users while monitoring key performance indicators before full release.
- **Federated Learning/Evaluation:** To leverage multi-institutional data without data sharing, federated learning allows training across sites. Similarly, federated evaluation frameworks would compute metrics locally and aggregate them, enabling cross-site monitoring even in competitive or privacy-sensitive contexts. For instance, a federated "AI registry" could anonymously pool AUC results from dozens of centers.
- **Explainability and Quality Assurance Tools:** Model interpretability methods (saliency maps, feature importance) can be used to audit whether an AI is focusing on reasonable features. Automated data quality dashboards (e.g. for image brightness, resolution, missing fields) can alert when an input data drift is occurring.
- **Clinical Decision Rules:** Some groups have proposed "out-of-distribution" detectors that flag when an input is too different from the training data (e.g. a new camera system or a rare pathology). If these detectors trigger frequently, it may signal the need for revalidation.
- **Synthetic Data and Simulation:** When real-world rare events are limited, synthetic data or in silico experiments (e.g. modifying images to simulate disease) might supplement evidence. However, synthetic data must be carefully validated to be credible.
- **Causal Inference Methods:** As noted in [49], causal modeling and counterfactual analysis may be needed to truly judge impact. For example, one can use instrumental variables (like random seed assignment of AI availability) or advanced statistical adjustment to estimate what would have happened without the AI. This area is nascent but promising.

All these technologies aim to make real-world evaluation feasible at scale, while preserving patient safety and compliance. But none replace well-designed clinical studies – they mainly augment surveillance and provide early warning of issues.

Implications and Future Directions

For Developers

AI SaMD manufacturers must plan for extensive evidence generation beyond initial testing. Engaging early with regulators and clinical partners is advisable. Developers should document the clinical rationale and intended use narrowly to guide study design. Building post-market monitoring into the product lifecycle (a requirement by IMDRF principles) is critical. Incremental improvements should be accompanied by defined analytic plans, in line with the FDA's proposed total life cycle guidance ([15] www.fda.gov).

Investment in data infrastructure pays dividends: for example, partnering with a health system to obtain diverse EHR data for validation, or collaborating on a multi-center trial. Transparency is becoming a competitive advantage: companies that register their studies and publish real-world results (even null findings) will build trust. In the long term, as suggested by G7's vision, good AI practice means publishing performance stratified by population subgroups (www.gov.uk), ensuring clinicians can choose appropriate use scenarios.

For Regulators

The traditional approval model is evolving. Regulators must develop clear pathways for ongoing oversight. This might include requiring manufacturers to submit periodic performance reports or mandating post-market studies (up to real-world Phase IV). The recent FDA RFI and draft guidance indicate such mechanisms are on the horizon. Collaboration across agencies (e.g. via IMDRF) to harmonize requirements will be important, as the G7 report recommends (www.gov.uk). Regulators should also provide clarity on what constitutes acceptable evidence: for example, is a large observational study enough to clear a new indication, or is an RCT needed? Guidance on statistical methods for real-world data will be helpful.

The FDA's AI-Enabled Medical Device list ([39] www.fda.gov) (a database of cleared AI devices) can serve as a resource, but regulators might also consider an accessible registry of unpublished performance data. Given the emphasis on transparency, revision of labeling to include information on the populations studied (age range, race/ethnicity, device versions) may be warranted.

For Clinicians and Health Systems

End-users should demand evidence appropriate to their setting. A hospital implementing an AI tool should ideally pilot-test it (e.g. running parallel to normal care for a period) to confirm performance in their patient population. Clinicians need training to understand the strengths and limitations of the AI, especially regarding when it might fail (e.g. new diseases, image artifacts). Workflow integration is non-trivial; evidence from workflow studies (like the mammography one) can guide institutions on resource allocation (e.g. need for additional readers).

Health systems can contribute to safety by tracking outcomes. For instance, if an AI tool misses a diagnosis, that should be captured as a safety event. Systems with robust registries or quality programs are well-positioned to evaluate AI impact (e.g. have AI-augmented colonoscopy outcomes as a metric in GI quality improvement).

For Patients

From the patient perspective, the bottom-line is whether AI improves care or safety. Well-designed real-world studies should address outcomes meaningful to patients: does AI lead to earlier cancer detection, less invasive interventions, more accurate diagnoses? Communicating the level of evidence underpinning an AI tool can help patients make informed choices. Ethical considerations – privacy, algorithmic fairness – should be transparently managed. Healthcare organizations may need to explain how patient data are used to train and evaluate AI models, fostering trust.

Conclusion

AI-based SaMD diagnostic tools are advancing medicine, but they also challenge traditional concepts of medical validation. This report has reviewed *how* to build credible clinical evidence for these tools, especially through **real-world performance studies**. We highlighted that regulators and experts now expect a continuum of evidence: from rigorous retrospective and prospective studies pre-market, to continuous monitoring and repeated evaluation post-market ([1] pmc.ncbi.nlm.nih.gov) ([2] www.fda.gov).

Key takeaways include:

- No single study type suffices. A mix of retrospective benchmarks, prospective trials, and real-world observational studies is needed.
- Performance must be measured in context. As the G7 panel stated, evaluations should occur in relevant clinical settings, with eye toward minimizing patient risk and workflow disruption (www.gov.uk).

- Metrics should be patient-centered and ethically robust. Use of reporting standards (CONSORT-AI, STARD-AI) and analysis by subgroups fosters transparency (www.gov.uk) (www.gov.uk).
- Post-market vigilance is crucial. Ongoing surveillance of AI performance, with predefined triggers for action, is emerging as best practice ([2] www.fda.gov) (www.gov.uk).

To realize the promise of AI diagnostics, stakeholders must commit to building this evidence infrastructure. Regulators should continue to provide clear guidance; researchers and companies should invest in large-scale multi-site studies and data-sharing consortia; clinicians should contribute to and utilize real-world data; and interdisciplinary collaboration (computer science, clinical medicine, regulatory science, ethics) is essential. In the words of experts, only by ensuring that AI tools consistently "work in theory and truly deliver value in clinical practice" will patient care benefit safely from this new wave of technology ([23] www.nature.com).

References: (All statements supported by sources [10–63] as cited in-text)

External Sources

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:,with...

[2] https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device#:~:Curre...

[3] https://www.nature.com/articles/s41591-023-02625-9#:~:imple...

[4] https://www.nature.com/articles/s41746-018-0040-6#:~:of%20...

[5] https://www.mdpi.com/2077-0383/15/7/2609#:~:syste...

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC12380398/#:~:Predi...

[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC9994700/#:~:This%...

[8] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:regul...

[9] https://www.imdrf.org/working-groups/software-medical-device#:~:Softw...

[10] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:In%20...

[11] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:In%20...

[12] https://www.mdpi.com/2077-0383/15/7/2609#:~:To%20...

[13] https://www.nature.com/articles/s41591-023-02625-9#:~:Moder...

[14] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:did%2...

[15] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing#:~:This%...

[16] https://www.sciencedirect.com/science/article/pii/S1546144021007407#:~:The%2...

[17] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:Durin...

[18] https://pmc.ncbi.nlm.nih.gov/articles/PMC11655112/#:~:Recen...

[19] https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device#:~:AI%20...

[20] https://www.nature.com/articles/s41591-023-02625-9#:~:Main...

[21] https://www.nature.com/articles/s41591-023-02625-9#:~:effec...

[22] https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device#:~:4,Pro...

[23] https://www.nature.com/collections/hcdeibadid#:~:%2A%2...

[24] https://pmc.ncbi.nlm.nih.gov/articles/PMC9994700/#:~:post,...

[25] https://pmc.ncbi.nlm.nih.gov/articles/PMC9994700/#:~:Three...

[26] https://pmc.ncbi.nlm.nih.gov/articles/PMC9994700/#:~:studi...

[27] https://www.nature.com/articles/s41591-023-02625-9#:~:A%20t...

[28] https://www.nature.com/articles/s41591-023-02625-9#:~:A%20t...

[29] https://www.nature.com/articles/s41591-023-02625-9#:~:The%2...

[30] https://www.nature.com/articles/s41591-023-02625-9#:~:Patie...

[31] https://www.nature.com/articles/s41591-023-02625-9#:~:The%2...

[32] https://www.nature.com/articles/s41591-023-02625-9#:~:of%20...

[33] https://www.mdpi.com/2077-0383/15/7/2609#:~:Backg...

[34] https://www.mdpi.com/2077-0383/15/7/2609#:~:syste...

[35] https://www.mdpi.com/2077-0383/15/7/2609#:~:ident...

[36] https://www.mdpi.com/2077-0383/15/7/2609#:~:four%...

[37] https://www.mdpi.com/2077-0383/15/7/2609#:~:gastr...

[38] https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device#:~:1...

[39] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices?aff_id=1314#:~:The%2...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.