AI SPECIALTY COLLABORATIVE

AI Tool Evaluation Guide

AMA AI Specialty Collaborative

The AMA AI Specialty Collaborative brings together 21 medical specialty societies and associations to ensure that physicians are shaping the future of AI in health care. Convened by the American Medical Association (AMA), this group works to center the physician voice in how AI is designed, developed, and integrated into clinical practice.

Developed in collaboration with:

American Medical Association

The AMA is the physician's powerful ally in patient care. As the only medical association that convenes 190+ state and specialty medical societies and other critical stakeholders, the AMA represents physicians with a unified voice to all key players in health care. The AMA leverages its strength by removing the obstacles that interfere with patient care, leading the charge to prevent chronic disease and confront public health crises, and driving the future of medicine to tackle the biggest challenges in health care.

For more information, visit ama-assn.org.

Manatt Health

Manatt Health integrates legal and consulting services to better meet the complex needs of clients across the health care system. Combining legal excellence, firsthand experience in shaping public policy, sophisticated strategy insight and deep analytic capabilities, we provide uniquely valuable professional services to the full range of health industry players. Our diverse team of more than 275 attorneys and consultants from Manatt, Phelps & Phillips, LLP, and its consulting subsidiary, Manatt Health Strategies, LLC, is passionate about helping our clients advance their business interests, fulfill their missions and lead health care into the future.

For more information, visit manatt.com/Health.

Special thanks to the following members of the AI Specialty Collaborative for their generous contributions, time, and expertise.

- Darlene King, MD
- Ivy Lee, MD, FAAD
- Josh Lesko, MD, FACEP
- T. Y. Alvin Liu, MD
- Srinivasan Suresh, MD, MBA, FAAP

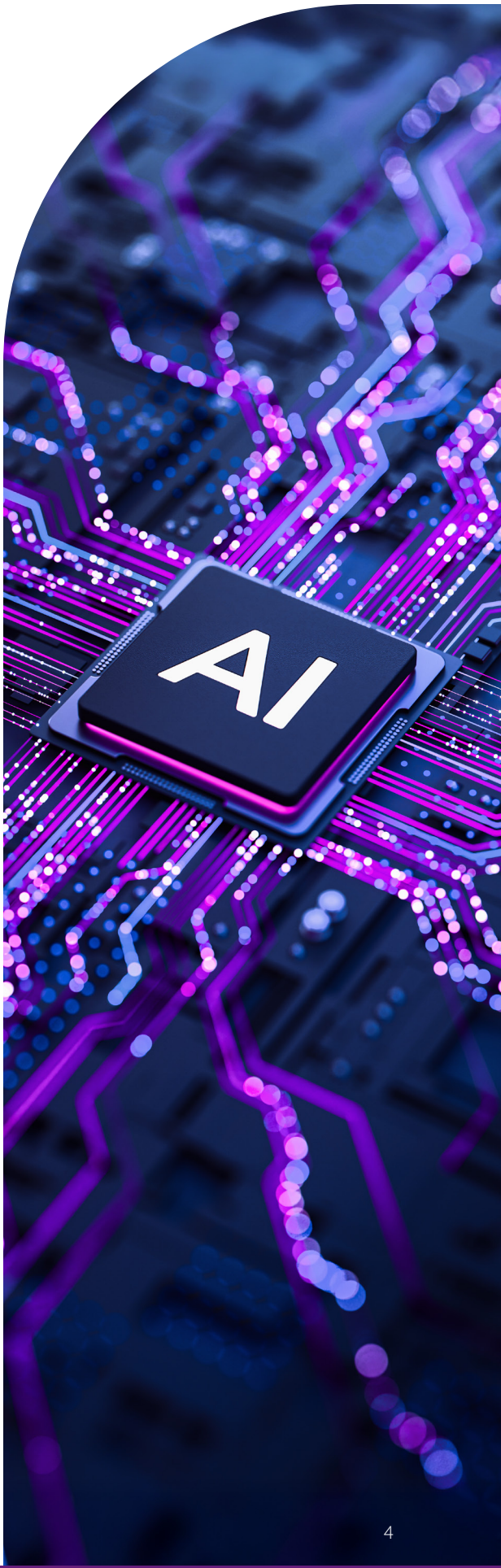
Background and context

With the growing availability of—and interest in—Augmented Intelligence (AI) applications in health care, the AMA is committed to ensuring physicians have the knowledge and tools to successfully navigate an AI-enabled future.

The AMA uses the term "augmented intelligence" in lieu of "artificial intelligence" to reflect its perspective that artificial intelligence tools and services support rather than explicitly replace human decision-making. For the purpose of this brief, AI refers to augmented intelligence.

With the rapid adoption of AI tools in medicine, it is essential that physicians are equipped to assess and manage the risks associated with their use in clinical practice. This resource, developed with physicians leaders from the AMA AI Specialty Collaborative, provides a structured framework for evaluating AI tools that support clinical decision-making, such as decision support tools, predictive models, and diagnostic or triage applications. It emphasizes clinical relevance, transparency, and patient safety and does not require deep technical expertise. This guide does not define *what* constitutes appropriate risk-taking in any given clinical setting. Instead, it outlines key questions to help physicians gather the information they need to determine what makes sense for their own practice.

This guide centers on understanding how to use AI "nutrition labels," or model cards, which are standardized summaries that describe an AI tool's intended purpose, data sources, performance metrics, limitations, and risks. Although designed to help physicians interpret model cards, the guide's five evaluation domains offer a broader framework for assessing AI tools, helping physicians interpret information, identify gaps, and engage vendors effectively.



How to use this guide

This guide supports physicians in assessing AI tools and provides developers with a clinician-informed framework to better align their products and resources with real-world clinical needs.

It is structured to help:

EVALUATE AI TOOLS

Evaluate available information about an AI tool to determine how well it aligns with clinical roles, patient populations, and workflows.

ENGAGE WITH VENDORS EFFECTIVELY

Ask targeted questions, identify missing information, and clarify capabilities early in the evaluation process.

ADAPT FOR SPECIALTY USE

Tailor evaluation to specialty-specific needs by integrating relevant considerations into education, reviews, and checklists.

ALIGN CROSS-FUNCTIONAL STAKEHOLDERS

Use a shared framework to help clinicians, IT, operations, and procurement teams reach consensus on performance, risks, and implementation requirements.

INFORM DEVELOPMENT

Provide developers and industry partners with insights to improve documentation, address risks and gaps, and better align products with clinical expectations.

The five domains of AI tool evaluation

This guide is structured around five domains that together provide a practical framework for evaluating AI tools in clinical decision-making.

01

CLINICAL USE CASE AND USER

Define purpose, target users, regulatory status, and appropriate clinical settings.

02

TRAINING AND VALIDATION DATA RELEVANCE

Assess whether the training and validation data reflect the patient population, clinical setting, and specialty.

03

RISKS AND MITIGATION

Identify safety and security concerns, known limitations, and available safeguards.

04

EFFECTIVENESS AND PERFORMANCE

Review reported metrics, validation methods, and real-world effectiveness.

05

WORKFLOW INTEGRATION AND MONITORING

Evaluate workflow fit, system integration, and ability to process for ongoing performance monitoring.

While each domain focuses on a distinct aspect of evaluation, key cross-cutting considerations apply across all five:

- **REPRESENTATION**

  Does the tool reflect your population and clinical setting, and does it appropriately disclose any potential biases or gaps?

- **TRANSPARENCY**

  Are the data sources, methodologies, and limitations clearly and comprehensively disclosed?

- **CLINICAL RELEVANCE**

  Does the tool integrate with workflows to support real-world decision making?

- **ONGOING SAFETY**

  Is there a defined plan for post-deployment monitoring, updates, and communication with users?

Collectively, these domains and cross-cutting considerations establish a structured and repeatable approach to AI tool evaluation, one that can be adapted to any specialty, practice setting, or procurement process.

Understanding this guide

For each of the five domains, this guide details Key Considerations, Areas to Watch, and Digging Deeper questions.

KEY CONSIDERATIONS

Baseline information that should be established early to clarify how the tool fits into clinical practice and whether it meets core requirements.

AREAS TO WATCH

Indicators that may not disqualify a tool, but suggest areas where further inquiry is needed to make an informed decision.

DIGGING DEEPER

Ask more detailed follow-up questions to help clarify nuances, surface risks, and test vendor transparency.

SPECIALTY EXAMPLES

Illustrative scenarios contributed by the AMA's AI Specialty Collaborative physicians that translate the framework into real-world practice. These examples illustrate how evaluation questions translate across specialties, helping clinicians connect general principles to the unique workflows, risks, and outcomes of their specialty.

01
Clinical use case and user

What is this tool designed to do, who should use it, in what clinical setting, and for what activity?

KEY CONSIDERATIONS

- Clarify the tool's clinical function (e.g., triage, diagnosis, summarization, treatment planning) and whether its output informs or automates a clinical action.
- Confirm the tool's regulatory status[2] (e.g., FDA-cleared, approved, or none), as this indicates the extent of external review it has undergone and helps set expectations for risk management, monitoring requirements, and liability. The FDA label is a good source of information to better understand technical details of the device.
- Clarify the tool's intended setting (e.g., ambulatory, inpatient, intensive care, emergency room).
- Verify whether the intended user (e.g., physician, nurse, advanced practice provider, technician) matches the intended clinical user(s), and whether the tool assumes a particular level of training, experience, or clinical/technical specialization.

DIGGING DEEPER

- What imaging source(s) does the model require (CT, MRI, X-ray, ultrasound, etc.), and what are the technical specifications?
- What assumptions does the tool make about data availability, integrations, or clinician capacity to use outputs, and do these align with my practice setting?

AREAS TO WATCH

- Intended use or user is described vaguely or overly broadly (e.g., "for clinicians in the ambulatory setting").
- No exclusions or limitations are specified, especially for populations, clinical settings, or edge cases where performance has not been evaluated.
- Assumptions about workflow, technology, or available expertise are not clearly stated, making it difficult to know if the tool fits your practice environment.

SPECIALTY EXAMPLES

- **Dermatology:**
  For imaging tools used to diagnose skin cancer, confirm whether the tool is designed for use on all skin sites, such as mucosa, scalp, hands.

- **Primary Care and Family Medicine:**
  If a primary care practice uses a tool that stratifies risk of patients for hospitalization or other health care interventions, it is essential to understand the population for which the tool was designed. For example, was it trained on data of older adults with multiple chronic conditions, pediatrics, or a broader population? Was the training data from individuals seen in the ambulatory, emergency, or inpatient setting? Applying the tool beyond the population or care setting it was designed for may produce biased risk assessments, causing clinicians to potentially overlook complications or misallocate resources.

[2] See appendix for definitions of regulatory statuses applicable to AI tools.

02
Training, testing, and validation data relevance

Does the data used to develop and test this tool represent my patient population and practice environment?

KEY CONSIDERATIONS

- Understand the sources of the training and validation data (e.g., academic medical centers, community clinics, international centers, urban emergency rooms, ambulatory surgical centers) and whether it is relevant.
- Assess whether key patient demographics (e.g., age, race, sex, geography, or acuity) are appropriately represented and if the characteristics are relevant to the local deployment site.
- Identify what "ground truth" or clinical reference standard was used (e.g., biopsy confirmation, lab test, expert panel) and whether the labeling standards match clinical norms in your specialty.

DIGGING DEEPER

- Was validation performed on more recent data to confirm the model performs as intended over time (not just on historical datasets)?
- Look beyond overall dataset size. How many unique patients were included? How many data points per patient? What was the mix of normal vs. abnormal cases?

AREAS TO WATCH

- Data sources are not specified, making it unclear whether they reflect your practice environment.
- "Validated" is mentioned without details on how (e.g., no distinction between internal, external, or prospective validation). Without this detail, it is difficult to judge the strength of the evidence.
- Dataset composition is vague, such as no breakdown of unique patients, repeat data points, or case mix (e.g., normal vs. abnormal readings)

SPECIALTY EXAMPLES

- **Neurology:**
  Assess whether stroke models were trained on datasets that include both acute and subacute stroke presentations to ensure that the model can recognize a wide spectrum of stroke cases (from early, emergency presentation to later evaluation stages). It is also important that the model has been trained on data from different clinical environments, including community emergency departments and tertiary stroke centers[3].
- **Medical Oncology:**
  Specify the observation date interval over which the medical tool predictions were validated. Specify the observed time between capture of multiple data modes in the case of multi-modal data.

[3] Stroke is used here as an illustrative example within Neurology. There are other examples of neurology-specific guidance available, including through the American Academy of Neurology (AAN).

03
Risks and mitigation

What are the potential risks of using this tool and in what scenarios or populations, and how are those risks managed or reduced?

KEY CONSIDERATIONS

- Identify known limitations and scenarios in which the model is not intended to be used (e.g., pediatrics, rare conditions, intensive care) and assess whether those exclusions align with your patient population and practice.
- Require the AI developer to provide its data security and privacy policies and evaluate any limitations.
- Determine who is responsible for reviewing the tool's outputs and acting on them, and what level of clinical oversight or intervention is required to ensure safe use.

DIGGING DEEPER

- For generative AI based tools, has the tool been assessed for accuracy, and are there scenarios where this risk of incorrect or fabricated information is higher?
- What processes are in place to detect and respond to safety issues or adverse events once the tool is deployed?

AREAS TO WATCH

- Claims of "no notable biases" without subgroup performance data to support the claim.
- Lack of a clear statement on required clinical oversight, responsibility, or limitations of use.
- No information on known failure modes (e.g., when the tool is more likely to misclassify or fabricate information).
- No discussion of risk mitigation strategies (e.g., alerts, confidence scores, guardrails, human-in-the-loop review).

SPECIALTY EXAMPLES

- **Allergy/Immunology:**
  Documentation should identify scenarios where overlapping conditions may cause diagnostic error, such as when symptoms mimic or confound one another. For example, in tools addressing chronic rhinosinusitis with nasal polyps (CRSwNP), overlapping presentations with allergic rhinitis or asthma may obscure accurate diagnosis, highlighting the need to flag these areas of uncertainty.

- **Psychiatry:**
  Understand what monitoring, processes, and procedures are in place when users interact with chatbot or tools in unforeseen ways such as revealing suicidal ideation. Confirm that clinician involvement and testing was involved in developing guardrails.

04
Effectiveness and performance

How well does this tool perform, especially in settings like mine?

KEY CONSIDERATIONS

- Identify which performance metrics are reported (e.g., sensitivity, specificity, area under the receiver operator curve) and assess whether they are sufficient for the tool's intended function.
- Determine whether performance was validated in real-world or prospective settings or only on retrospective datasets. Ask for real-world studies if available.
- Clarify which clinical endpoints or outcomes were used to evaluate model performance (e.g., diagnostic accuracy, reduction in time to diagnosis, impact on workflow), and whether they align with your goals for deployment.

DIGGING DEEPER

- How does the tool's performance vary by clinical setting (e.g., community vs. academic, inpatient vs. outpatient)?
- How does the tool's reported performance compare with current standard of care or alternative approaches?
- To what extent can clinicians adjust or customize the tool's underlying clinical logic to reflect current clinical guidelines or local practice standards?

AREAS TO WATCH

- Tools reporting only retrospective performance without real-world validation.
- Overemphasis on single headline metrics (e.g., accuracy) without clarity on tradeoffs between sensitivity and specificity.
- Lack of subgroup analyses that may reveal poor performance in certain populations.

SPECIALTY EXAMPLES

- **Emergency Medicine:**
  Triage tools should emphasize sensitivity, given the risk of under-triaging and missing critical cases in this setting. While high sensitivity could come at the risk of lower specificity (e.g., false alarms), this may be acceptable in the emergency setting.

- **Nuclear Cardiology:**
  To ensure relevance and clinical utility, the accuracy of nuclear cardiology diagnostic tools for evaluating epicardial coronary artery disease should be benchmarked against gold standards, such as invasive coronary angiography (ICA) and fractional flow reserve (FFR). These comparisons establish a tool's ability to detect disease. The effectiveness of the tool also depends on its capacity to predict risk for coronary-related events, and cardiac-specific mortality.

05 Workflow integration and monitoring

How will this tool fit into my workflow, and how will its performance be monitored over time?

KEY CONSIDERATIONS

- Clarify how the tool integrates into existing clinical systems (e.g., Electronic Health Record (EHR), Picture Archiving and Communication System (PACS), or a standalone interface) and evaluate whether it streamlines workflows or introduces friction that could slow down care delivery.
- Identify how the tool's performance will be monitored after deployment (e.g., audit logs, feedback loops, drift tracking) and specify whether responsibility for this monitoring lies primarily with the vendor, the client organization, or a shared model.
- Determine whether the vendor has a defined process for updating the model, including how updates will be communicated, reviewed, and approved by the provider organization, and how operational issues such as downtime or re-training of staff will be managed.

DIGGING DEEPER

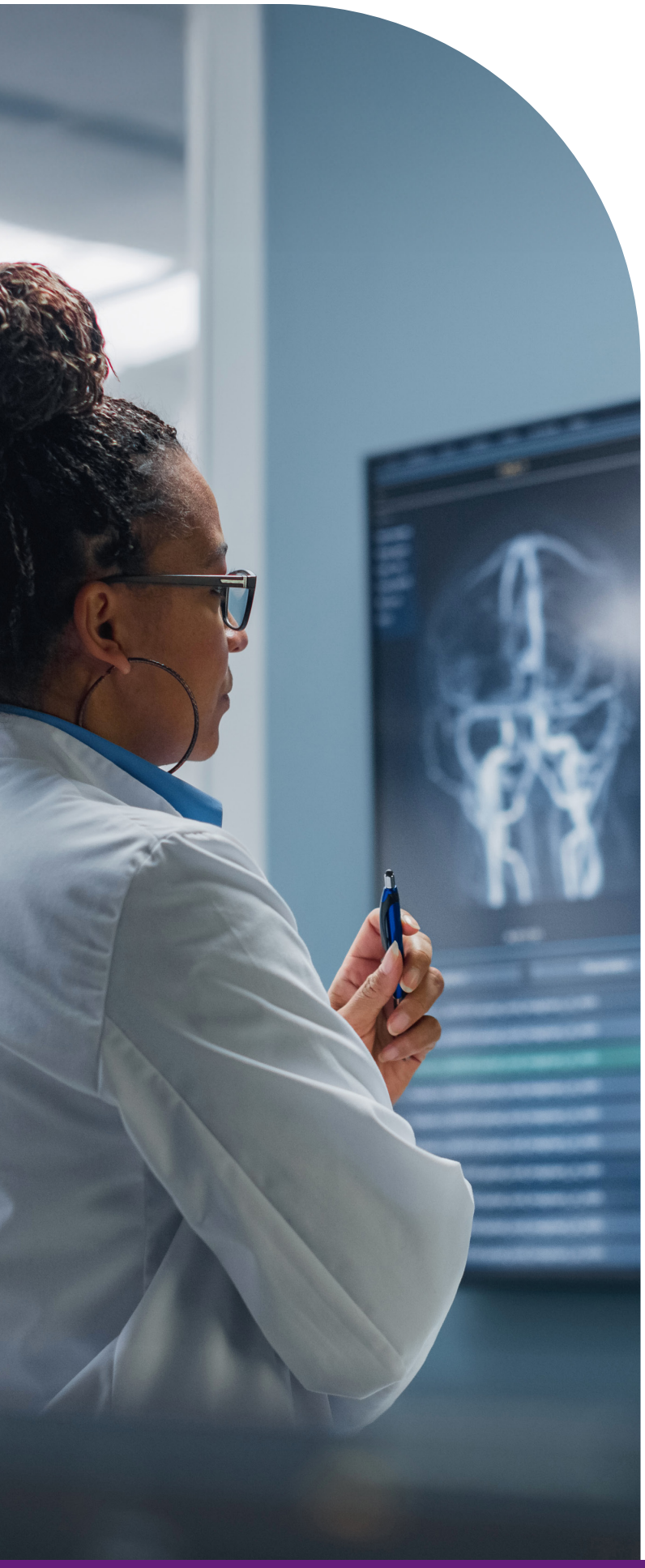
- Are the tool's outputs real-time, delayed, or batched, and how would that timing affect time-sensitive clinical decisions?
- Who within the organization (vendor versus client) is accountable for monitoring tool performance and responding to issues?
- Has the tool been tested for impact on clinician workload and cognitive burden during real-world use?

AREAS TO WATCH

- No mention of how model updates, retraining, or version control will be managed, or how changes will be communicated to clinicians.
- Tool requires use of a separate interface but does not describe training, onboarding, or support for adoption.
- Lack of clarity on performance monitoring post-deployment (e.g., who is responsible for drift detection, error reporting, or continuous assurance.
- Non-specific explanation of how downtime, failures, or integration errors (e.g., with EHR/PACS) will be handled in clinical workflows.

SPECIALTY EXAMPLES

- **Radiology:**
  AI-generated triage signals in imaging, such as those for suspected intracranial hemorrhage (ICH) or large vessel occlusion (LVO), can interact with other applications to help inform the priorities on a radiologist's worklist, trigger mobile alerts to stroke or emergency department teams, populate dedicated triage dashboards, or initiate tasks within the EHR.



SPECIALTY EXAMPLES CONTINUED

Therefore, it is important for organizations to clearly document which systems receive the triage signal, how priorities are established (such as queue positions, alerts, and timers), and who is responsible for follow-up actions. Standalone computer-aided detection triage (CADt) tools, the most prevalent radiology AI CAD tools in the U.S. market today, are designed primarily for notification and queuing purposes. By definition, such tools do not display AI findings on images, as image highlighting is a function of computer-aided detection (CADe) and diagnostic (CADx) tools. Standalone CADt user interfaces are limited to case status, timestamps, and routing logic rather than visual overlays. This reinforces the understanding of regulators like FDA that the output from triage is not a diagnostic result, and the images must be reviewed by the appropriate physician readers.

- **OB/GYN:**
  Tools that support labor and delivery monitoring should clearly state whether they integrate directly with existing fetal monitoring systems. Decision-makers should understand how the tool generates and prioritizes alerts for obstetric teams, so it is clear who is notified and when. They should also understand how the system manages downtime or technical failures during high-acuity events, when uninterrupted monitoring is most critical.

Regulatory pathways for AI tools in health care[4]

AI-enabled health care tools fall under different Food and Drug Administration medical device regulatory categories depending on their intended use, risk level, and claims. The main U.S. pathways are:

FDA-CLEARED

- **Pathway:** Premarket Notification 510(k) pathway
- **Description:** Intended for low- to moderate-risk devices. Clearance is granted when a tool is shown to be "substantially equivalent" to a legally marketed predicate device.
- **Implications for physicians:** Although there is a premarket notification submission, there is moderate oversight by the FDA, and clinical testing is not always necessary. It is likely safe but careful due diligence is required by adopting physician.
- **Example:** Aidoc BriefCase for Intracranial Hemorrhage

510(K) EXEMPT

- **Pathway:** Regulated as a medical device, but no premarket submission is required.
- **Description:** Intended for low- to moderate-risk devices that have been exempted from premarket notification requirements by regulation.
- **Implications for physicians:** Certain regulatory controls still apply, but FDA does not perform any review of the product before it is distributed. FDA has deemed these products to be lower risk, but adopting physicians should exercise care to identify reputable suppliers.
- **Example:** Continuous glucose monitor retrospective data analysis software.

[4] Artificial Intelligence in Software as a Medical Device

FDA DE NOVO AUTHORIZED

- **Pathway:** De Novo Classification Request
- **Description:** Intended for novel, low- to moderate-risk (Class I or II) devices without an existing predicate device. A granted De Novo request establishes a new device type and special controls that may be referenced by subsequent 510(k) submissions.
- **Implications for physicians:** De Novo submissions are typically more comprehensive and include more data than 510(k) submissions. However, since there isn't precedent in the market, physicians should monitor performance closely.
- **Example:** IDx-DR (Digital Diagnostics) - A first-of-its-kind AI tool for early detection of diabetic retinopathy.

Note: *510(k) and De Novo Authorization pathways are currently used for some prescription digital therapeutic (PDTx) products. These pathways are designed for medical devices and do not have equivalent evidence requirements as the pathways used for traditional medications.*

FDA-APPROVED

- **Pathway:** Premarket Approval
- **Description:** Required for high-risk (Class III) devices that support or sustain human life or present a potential unreasonable risk of illness or injury. These devices must demonstrate reasonable assurance of safety and effectiveness through valid scientific evidence, typically clinical trials. The least common pathway for AI but relevant for tools making direct treatment decisions.
- **Implications for physicians:** These are typically reserved for higher risk use cases but given the level of evidence and clinical testing, typically reliable for use as they were approved.
- **Example:** An AI-driven device that autonomously diagnoses and recommends treatment without clinician review.

NOT FDA-REGULATED

- **Description:** Software intended for medical purposes that falls within an exemption under the 21st Century Cures Act.
- **Example:** Clinical decision support (CDS) software tools that provide recommendations to a health care professional, enable the professional to independently review the basis for the recommendation, but are not intended to replace clinical judgement.

Understanding an AI tool's regulatory journey is an important component of evaluating a tool, as classification and perceived level of risk can help a physician.

Glossary of terms

Augmented Intelligence:
A conceptualization of artificial intelligence that emphasizes its assistive role, designed to enhance, not replace, human intelligence and clinical decision-making.

Bias:
Systematic errors in data or models that can lead to unfair or inaccurate outcomes. Bias may arise from training data (e.g., sample bias, measurement bias) or human behavior (e.g., reliance bias).

Confabulation:
An AI-generated output that is factually incorrect, misleading, or nonsensical, often presented as if it were accurate. May also be referred to as hallucination or fabrication.

Foundation models:
Large-scale AI models trained on vast and diverse datasets that can be adapted to many downstream tasks (e.g., large language models like GPT).

Generative AI (GenAI):
AI systems capable of producing new content (e.g., text, images, etc.) based on learned patterns from training data. Examples include chatbots and text summarization tools.

Large Language Model (LLM):
A type of generative AI trained to understand and generate human language, often used in applications like chatbots or clinical documentation.

Machine Learning:
A subtype of AI where algorithms learn from data to make predictions or decisions without being explicitly programmed for each task.

Natural Language Processing (NLP):
A branch of AI focused on enabling computers to understand, interpret, and generate human language.

Supervised Learning:
A machine learning approach in which models are trained on labeled data, learning to map inputs to known outputs.

Testing data:
A separate and final dataset used after training and validation to objectively assess a model's real-world performance. It provides an unbiased evaluation of how the model will perform on unseen data and is often used for reporting performance metrics.

Unsupervised Learning:
A machine learning method where algorithms analyze unlabeled data to find hidden patterns or groupings without pre-defined outcomes.

Validation data:
A subset of data used during model development to tune parameters and assess model performance after training.