



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

FDA AI Credibility Assessment in Drug Development

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 · 2026-03-04 · fda ai framework · ai credibility assessment · drug development · ai model validation · regulatory compliance · context of use · machine learning · risk-based validation

Original article: <https://intuitionlabs.ai/articles/fda-ai-credibility-assessment-drug-development>



In this article

1. Executive Summary
 2. Introduction and Background
 3. FDA's AI Credibility Framework: Defining Context and Risk
 4. Step 1: Define the Question of Interest
 5. Step 2: Define the Context of Use (COU)
 6. Step 3: Assess Model Risk (Influence and Consequence)
 7. Development and Validation of AI Models in Drug Development
 8. Data Quality and Fitness for Use
 9. Model Development and Architecture
 10. Performance Validation (Steps 4–5)
 11. Cross-Cutting Considerations: Uncertainty, Bias, and Explainability
 12. Step 4–7: Planning, Execution, Documentation, and Review
 13. Data Analysis and Evidence – Building Credibility
 14. Ensuring High-Quality Training Data
 15. Performance Evaluation
 16. Dealing with Bias and Fairness
 17. Documentation of Analysis
 18. Real-World Case Studies
 19. Case 1: AI in Bioprocess Optimization (Ark Biotech)
 20. Case 2: AI for Patient Selection (COVID-19 Trial)
 21. Case 3: AI-Enabled Image Analysis for Endpoints (NASH Trial)
 22. Documentation and Regulatory Submission
 23. Implications, Challenges, and Future Directions
 24. Conclusion
-

Executive Summary

The U.S. Food and Drug Administration (FDA) has recently introduced a **risk-based credibility assessment framework** for AI models used in drug and biological product development ([regulations.justia.com](https://www.regulations.gov)) ([emergingaihub.com](https://www.emergingaihub.com)). In January 2025 the FDA released draft guidance (*Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*) that outlines **seven steps** sponsors should follow to ensure any AI/ML model's output is trustworthy for its intended regulatory context ([regulations.justia.com](https://www.regulations.gov)) ([emergingaihub.com](https://www.emergingaihub.com)). The cornerstone of the framework is *context of use (COU)* – i.e. the precise question or decision the model supports – and *credibility evidence*, meaning the data, tests, and documentation needed to establish trust in the model's performance for that COU ([regulations.justia.com](https://www.regulations.gov)) ([emergingaihub.com](https://www.emergingaihub.com)).

A **risk-based approach** pervades the proposed framework. The draft recommends tailoring credibility activities to model risk and context of use. It does not create a blanket requirement to submit every model's full architecture, training data and uncertainty analysis; the information FDA may request depends on risk and

COU ([FDA draft guidance, section IV.A.4](#)).

Key components of evaluation include ensuring **high-quality, “fit-for-use” data**, robust **model validation** on independent datasets, and comprehensive performance metrics (accuracy, confidence intervals, bias measures, etc.) ([pmc.ncbi.nlm.nih.gov](#)) ([emergingaihub.com](#)). The guidance emphasizes documenting all activities: defining the regulatory question (Step 1), specifying COU (Step 2), performing a risk analysis (Step 3), planning and executing validation studies (Steps 4–5), and compiling results and deviations (Step 6) ([regulations.justia.com](#)) ([emergingaihub.com](#)). Finally, Step 7 of the proposed framework addresses whether the evidence establishes adequate credibility for the COU ([FDA draft guidance](#)). Throughout, transparency is stressed: FDA expects clear reporting of model development, testing, and changes so reviewers can **independently evaluate** the AI evidence without rerunning analyses ([emergingaihub.com](#)) ([www.foley.com](#)).

This report provides an in-depth treatment of the FDA's AI credibility framework. We first present historical context on AI/ML in drug development and prior regulatory initiatives. We then unpack each of the seven steps with detailed analysis and expert commentary, covering how to define questions, set COUs, assess risk (influence × consequence), and plan validation. Sections on **data quality, model validation, and performance evaluation** delve into the technical evidence (data curation, bias analysis, uncertainty quantification) needed for credibility. We also discuss **documentation practices**, including what sponsors should include in submissions (data provenance, code descriptions, [audit trails](#), [Part 11 compliance](#)) and how to balance transparency with intellectual property concerns ([www.foley.com](#)).

The examples include Ark's illustrative monoclonal-antibody process-development scenario ([Ark Biotech](#)). Separately, on November 8, 2022, FDA issued an EUA for Kineret (anakinra) for COVID-19 in hospitalized adults with positive direct SARS-CoV-2 tests and pneumonia requiring supplemental oxygen, at risk of severe respiratory failure and likely to have elevated suPAR. SCORE2 supported identification of that population ([FDA review](#)). EMA announced AIM-NASH's qualification in March 2025; its defined use assists pathologists in phase 2 and 3 MASH trials ([EMA qualification opinion](#)).

Finally, we explore broader implications and future directions. The FDA's emphasis on transparency and rigor is expected to spur innovations – for example, tools for explainable AI, automated report generation, and model monitoring ([www.foley.com](#)) ([FDA draft guidance, section IV.B](#)). Sponsors should prepare by engaging FDA early and aligning their validation plans with the model's risk level. We compare FDA's approach to emerging global guidelines (e.g. the EMA's reflection paper on AI in the pharmaceutical lifecycle) and note remaining challenges in harmonization ([www.ema.europa.eu](#)) ([emergingaihub.com](#)). Ultimately, the FDA's framework lays out a path for building *trustworthy AI* in drug development – offering nonbinding recommendations ([FDA draft guidance](#)) – so that AI can deliver its promise of innovation while upholding patient safety and product quality ([regulations.justia.com](#)) ([emergingaihub.com](#)).

<details><summary>Table: Seven Steps of FDA's AI Model Credibility Framework</summary>

Step	Name	Description (Regulatory Focus)	Key Guidances/Citations
1	Define Question of Interest	Articulate the specific <i>decision or outcome</i> the model is intended to inform (e.g. “identify trial-eligible patients” or “determine batch fill-volume compliance”). This ties the AI use to a concrete regulatory objective (safety, efficacy, quality).	E.g., define exact inclusion/exclusion criteria target (emergingaihub.com) (regulations.justia.com)
2	Define Context of Use (COU)	Specify <i>how</i> the model will be used: its role relative to other evidence, degree of automation vs. human oversight, and the conditions under which it operates. This sets boundaries for its safe application (e.g. as a support tool vs. an autonomous classifier).	E.g., “AI will flag patients for review by investigator” vs “AI makes final call” (emergingaihub.com) (www.jdsupra.com)
3	Assess Model Risk	Identify risks using an FDA-proposed matrix (model influence × decision consequence). Rate how much the model's output drives decisions and the potential impact if it errs. High-risk scenarios (high influence and/or high consequence) demand more scrutiny and mitigation.	Determine influence and consequence (emergingaihub.com) (regulations.justia.com)
4	Credibility Plan	Based on the risk, develop a validation plan outlining needed evidence. Include data quality checks, model development descriptions, choice of performance metrics, statistical validation approaches (with independent test sets), and bias/fairness assessments.	Plan should cover data fitness, performance metrics, validation strategy (emergingaihub.com)
5	Execute Plan	Conduct the planned validation and testing activities. This means running the model on test data, computing metrics (accuracy, error rates, confidence intervals, etc.), documenting results in real time. Ensure auditability (e.g. electronic records per 21 CFR Part 11 for review by FDA).	Perform model evaluations, collect metrics (pmc.ncbi.nlm.nih.gov) (emergingaihub.com)
6	Document Results	Compile a Credibility Assessment Report detailing all validation outcomes and any deviations from the plan. Include analysis of results, bias checks, error analysis, and justification for any anomalies. Documentation must be clear enough for FDA reviewers to assess credibility.	Include full results, deviations rationale (emergingaihub.com) (regulations.justia.com)
7	Determine Adequacy	Evaluate whether the accumulated evidence shows the model is sufficiently reliable for the COU. This is a judgment step: consider the model's performance relative to risk, note remaining uncertainties, and decide if it can be submitted. Plan for monitoring if COU or data shift.	Assess adequacy vs. risk, plan monitoring (lifecycle) (emergingaihub.com) (FDA draft guidance, section IV.B)

</details>

Introduction and Background

Artificial intelligence (AI) and machine learning (ML) are **transforming drug development**, spanning tasks from target identification to clinical trial design and manufacturing optimization. Sponsors increasingly use AI/ML to predict patient outcomes, identify trial cohorts, analyze biomarkers, and optimize bioprocesses. For example, ML models can integrate large real-world datasets to find patient subgroups or simulate “digital twin” trials ([pmc.ncbi.nlm.nih.gov](https://www.ncbi.nlm.nih.gov/pmc)). FDA Commissioner Califf has noted AI’s “transformative potential to advance clinical research” when proper safeguards are in place (www.fda.gov).

However, these advanced models also introduce new challenges. Many modern AI systems (e.g. deep neural networks or ensemble algorithms) are essentially “black boxes” whose internal logic is not easily interpretable ([academic.oup.com](#)). This opacity raises concerns: biases or errors in training data may be amplified, model predictions may not generalize across patient groups, and complex models may be difficult for regulators to verify ([academic.oup.com](#)) ([pmc.ncbi.nlm.nih.gov](#)). Given that drug development decisions directly affect patient safety and product quality, regulators must ensure AI tools are *reliable and valid* for their intended use.

The FDA has long encouraged innovation but has also needed to clarify how AI/ML fits into existing regulatory pathways. Over the past decade, the agency has seen a surge in submissions involving AI “components” ([www.fda.gov](#)) ([regulations.justia.com](#)). For instance, since 2016 FDA centers have reviewed on the order of 300–1,000 drug- or biologics-related submissions incorporating AI/ML methods ([www.fda.gov](#)) ([emergingaihub.com](#)). Common applications have included predictive models for clinical trial enrollment, digital biomarkers, and advanced analytics for manufacturing and pharmacovigilance. In tandem, stakeholders (industry, academia, patients) have pushed FDA for guidance on how AI should be evaluated. FDA has organized workshops (e.g. with Duke’s Margolis Center in 2022) and published discussion documents in 2023 (soliciting **800+ comments**) on AI in drug development and manufacturing ([www.fda.gov](#)) ([regulations.justia.com](#)).

In January 2025, the FDA issued its first draft guidance focused on *drug development* AI (complementing a 2023 guidance on AI-enabled medical devices) ([www.fda.gov](#)). Titled “*Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*”, the draft guidance provides recommendations on how sponsors should plan, validate, and document AI models across the drug lifecycle (nonclinical, clinical, manufacturing) when those models produce data or analyses that will inform regulatory decisions ([regulations.justia.com](#)) ([www.dlapiper.com](#)). The draft defines *context of use* (COU) as the specific role and scope of the model in addressing a question of interest ([FDA draft guidance](#)) ([regulations.justia.com](#)) ([emergingaihub.com](#)). Crucially, the guidance introduces a **seven-step credibility assessment** that sponsors can follow to demonstrate trust in a model’s output. This framework upholds FDA’s traditional standards for evidence, framing them as *credibility evidence* for AI–ML models ([regulations.justia.com](#)) ([emergingaihub.com](#)).

Notably, the draft guidance explicitly **excludes** early drug discovery (e.g. target ID, lead optimization) and purely internal uses (workflow tools, administrative tasks) that do not impact patient safety or product quality ([regulations.justia.com](#)) ([emergingaihub.com](#)). In other words, it applies only when “AI-generated output is intended to support a regulatory decision” about a drug’s quality, safety, or efficacy ([regulations.justia.com](#)). For example, an ML model that flags adverse events or predicts trial outcomes falls within scope, whereas an AI tool used only to design molecules (before any submission) does not ([regulations.justia.com](#)) ([emergingaihub.com](#)). By focusing on decision-impacting AI, the guidance seeks to ensure that high-stakes uses of AI are subject to rigorous review.

The FDA’s draft emphasizes **transparency and documentation**. It describes “credibility” as “*trust, established through the collection of credibility evidence, in the performance of an AI model for a particular [COU]*” ([regulations.justia.com](#)). Credibility evidence can be any data or analysis that supports the model’s validity (training data details, validation results, etc.). The agency encourages early communication with FDA staff to

agree on appropriate credibility activities based on model risk ([regulations.justia.com](https://www.regulations.gov)). As the draft states, from COU to documentation, the proposed recommendations “will generally be tailored to the specific COU and will depend on model risk” ([regulations.justia.com](https://www.regulations.gov)).

This research report draws on the FDA’s text and public statements (www.fda.gov) ([regulations.justia.com](https://www.regulations.gov)), expert commentaries (www.foley.com) (www.dlapiper.com), academic analyses (academic.oup.com) ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)), and case examples to unpack the FDA’s AI guidance. We detail each of the seven credibility steps, discuss data and validation considerations (e.g. metrics, bias, uncertainty), and review how sponsors should document their AI work. We also survey multiple perspectives—from legal experts, industry practitioners, and international regulators—on opportunities and challenges. The examples distinguish an illustrative monoclonal-antibody process-development scenario ([Ark Biotech](https://www.arkbiotech.com)) from documented regulatory analyses. By the end, readers will have a comprehensive view of how to evaluate an AI model’s credibility in the context of drug development, how to assemble the required evidence, and what this evolving regulatory landscape implies for the innovation of AI in pharmaceuticals.

FDA’s AI Credibility Framework: Defining Context and Risk

The heart of the FDA’s draft guidance is a **risk-based credibility framework** for AI models that influence drug development decisions ([regulations.justia.com](https://www.regulations.gov)) (www.dlapiper.com). This framework revolves around two foundational concepts:

- **Context of Use (COU).** The COU precisely defines *how* and *where* the AI model will be applied to a specific decision. It includes the model’s scope, inputs and outputs, and its role in the workflow (e.g. a screening tool vs. an automated classifier). For example, a COU might be “using an ML model on EHR data to predict which trial candidates meet inclusion criteria, to inform investigator referrals” (flagging with human review), versus “AI automatically assigns batch-release clearance in manufacturing without human check” ([emergingaihub.com](https://www.emergingaihub.com)) (www.jdsupra.com). Clarifying COU ensures the model is evaluated *only for its intended task*, not generalized beyond it ([regulations.justia.com](https://www.regulations.gov)) ([emergingaihub.com](https://www.emergingaihub.com)).
- **Credibility Evidence.** The proposed framework supports establishing and evaluating model credibility for a particular COU ([FDA guidance page](https://www.fda.gov)). The guidance defines credibility as “trust...established through the collection of credibility evidence, in the performance of an AI model for a particular COU” ([regulations.justia.com](https://www.regulations.gov)). In practice, this means documenting the model design, training, and validation so that reviewers can independently assess its reliability. Credibility evidence can include training data descriptions, performance metrics (accuracy, AUC, calibration), statistical uncertainty (confidence intervals, error rates), validation on hold-out data, results of robustness tests, etc.—any data relevant to making the model’s output believable for the COU ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ([emergingaihub.com](https://www.emergingaihub.com)).

The FDA’s proposed credibility approach is **risk-based**. Model influence concerns the contribution of model-derived evidence relative to other evidence; decision consequence concerns the significance of an adverse outcome from an incorrect decision. Human override alone does not establish low influence or low overall risk.

Assess reliance, corroborating evidence and decision consequences in the specific COU ([FDA draft guidance, section IV.A.3](#)).

In practical terms, most uses within the guidance's scope will tend toward high-risk. As noted by Foley LLP, AI used for trial management or manufacturing likely *influences* critical decisions, so sponsors should plan to submit extensive details about the AI model ([www.jdsupra.com](#)). Indeed, the draft guidance emphasizes that high-risk models are expected to include **comprehensive disclosures**: the model's architecture, full description of training and source data, algorithms, validation methods, and performance statistics ([www.jdsupra.com](#)) ([pmc.ncbi.nlm.nih.gov](#)). Lower-risk models might allow pared-down submission (e.g. summary performance). But in either case, the principle is the same: the higher the potential impact, the more rigor needed in building and documenting trust.

Below we outline the first key steps in setting up an AI credibility strategy: defining the question and context, and assessing risk. Later sections delve into the validation and documentation needed to address that risk.

Step 1: Define the Question of Interest

The **first step** is to articulate the precise “question of interest” the AI model is meant to answer ([emergingaihub.com](#)) ([www.dlapiper.com](#)). This is essentially a problem statement tied to the FDA's review goals (safety, efficacy, quality). It forces sponsors to ground their AI use-case in a concrete regulatory decision rather than a vague business goal. For instance:

- In a clinical trial setting, the question might be: “Which patients in trial XYZ are eligible for outpatient monitoring after dosing Drug A, without compromising safety?” (This implicates patient safety, since missing an adverse event could be serious.)
- In manufacturing, a question could be: “Can batch Y, produced under process conditions Z, be released as meeting fill-volume specifications?” (This affects product quality.)

Defining the question is critical because it determines both the *inputs/outputs* of the model and the *acceptance criteria*. As Foley LLP observes, examples include using AI for *trial inclusion criteria*, *patient risk stratification*, or *endpoint adjudication* ([www.dlapiper.com](#)). Sponsors should describe how the AI output and other evidence contribute to the decision. Human review alone does not establish lower risk; assess model influence and decision consequence for the actual COU ([FDA draft guidance](#)).

In practice, sponsors should document the question of interest early and include any supporting context or historical data. For example, if a model predicts patient dropout risk, one might point to prior trial data on dropout rates. The question then shapes the context of use, model design, and evaluation plan. As the Emerging AI Hub notes, framing the question precisely influences downstream planning: “the specificity matters because everything downstream flows from how you’ve framed this question” ([emergingaihub.com](#)). Without a clear question, validation metrics and risk analysis cannot be properly defined.

Step 2: Define the Context of Use (COU)

Once the question is set, **Step 2** is to define the **Context of Use (COU)**: how exactly the AI model will operate to address that question ([emergingaihub.com](https://www.emergingaihub.com)) ([regulations.justia.com](https://www.regulations.justia.com)). The COU adds the operational details around the reference question. It answers: What data will the model use? At what point in the process? Who sees the output and does what with it? For example, a COU description might specify: “The AI model will process ECG and lab data for each screening patient, output a binary flag of ‘eligible’ or ‘not eligible’, and an investigator will be notified to make the inclusion decision.”

The COU component is essential for scoping the evaluation. It includes:

- **Role of AI:** Is the model a decision-support tool or an automatic classifier? Will its output be reviewed by a human, or will it feed directly into an automated control system?
- **Boundaries and Inputs:** What inputs and training data are permissible? What patient population or batch conditions?
- **Integration with other evidence:** Is the AI the sole evidence for a decision, or one piece among many?

FDA guidance and commentators stress that understanding the AI's role (particularly the *degree of autonomy*) is crucial. For example, if an AI simply screens patients and a doctor makes the final call, the risk picture differs from an AI that autonomously enrolls or disqualifies patients. FDA's draft defines COU as “the specific role and scope of the AI model used to address a question of interest” ([FDA draft guidance](#)). Accurately describing COU provides the baseline for all subsequent risk assessment and validation.

Several illustrative contexts of use are mentioned in the guidance, including:

- **Clinical Trial Design and Management** (e.g. optimizing inclusion criteria)
- **Patient Evaluation** (e.g. medical imaging or biomarker interpretation)
- **Endpoint Adjudication** (e.g. applying a digital biomarker to determine a clinical outcome)
- **Analyzing Trial Data** (e.g. meta-analytic or signal detection tools)
- **Pharmacovigilance** (e.g. AI triage of adverse event reports)
- **Manufacturing and Quality Control** (e.g. ensuring fill-volume or sterility using sensors) (www.jdsupra.com).

The context description also flags uses outside the scope. The FDA guidance explicitly excludes flows such as early *discovery* (where molecules are generated by AI but then evaluated by traditional methods) and *operational* uses that do not directly affect safety or quality (e.g. scheduling or automated report writing) ([regulations.justia.com](https://www.regulations.justia.com)) ([emergingaihub.com](https://www.emergingaihub.com)). By focusing on COU, sponsors can confirm whether their AI use needs to follow these FDA recommendations (if it impacts regulatory decisions) or can use more proprietary development (e.g. trade secret models not submitted).

Step 3: Assess Model Risk (Influence and Consequence)

With the COU in hand, **Step 3** is to **assess the model's risk level**, combining *model influence* and *decision consequence* ([emergingaihub.com](https://www.emergingaihub.com)) ([regulations.justia.com](https://www.regulations.justia.com)). This determines how stringent the rest of the credibility activities must be. The FDA proposes a two-dimensional risk matrix:

- **Model Influence:** assess the contribution of model-derived evidence relative to other evidence used to answer the question. The ability to override an alert does not by itself demonstrate low influence ([FDA draft guidance](#)).
- **Decision Consequence:** how severe are the outcomes if the model is wrong? Does an error merely inconvenience the process, or does it risk patient harm or defective products? A misclassification of an insignificant lab value has low consequence; misidentifying a malignant tumor as benign (or failing to detect it) would be very high consequence.

The combination yields a risk rating (low, medium, high, etc.). A “high-risk” AI use might be high influence *and* high consequence. For example, the guidance gives a scenario where an AI assigns trial participants to home care vs. inpatient monitoring. If the model incorrectly deems a patient ‘low risk’ who is actually at high risk, the patient could suffer serious harm. This is a high-influence, high-consequence context, demanding intense scrutiny ([emergingaihub.com](#)). In contrast, an AI that, say, optimizes an administrative scheduling task (low consequence, low influence) would require minimal evidence.

Sponsors should **document their risk assessment rationale**. FDA notes that if the COU changes – e.g. the model is applied to a new patient population or the oversight differs – the risk analysis must be revisited ([emergingaihub.com](#)). The Emerging AI Hub emphasizes that risk assessment is not a one-time checkbox but a **living document**: assumptions about influence and consequence should be recorded and updated as needed ([emergingaihub.com](#)). In practice, this might take the form of a written justification or risk matrix table, explaining why each factor was scored as it was, so FDA reviewers can understand the sponsor’s judgment.

Once risk is classified, it sets the **scope of credibility activities**. A high-risk model will trigger deeper investigations (e.g., extensive testing across edge cases, stricter performance criteria) and require submission of details such as code and raw data if feasible ([www.jdsupra.com](#)) ([pmc.ncbi.nlm.nih.gov](#)). A lower-risk model might need only summary results and supporting descriptions. But in all cases, sponsors must show that the level of evidence is commensurate with their risk categorization and COU.

<details><summary>Table: Key Documentation and Validation by Risk Level (Summary)</summary>

Risk Level	Use-case Examples	Documentation/Validation Expected	Implications
High Risk	A COU with high reliance on model evidence and serious consequences from an incorrect decision	Credibility activities and performance acceptance criteria tailored to that COU and risk	More stringent assessment than for lower-risk COUs
Medium (conditional)	Safety-signal review or manufacturing recommendations only where the specified COU, reliance on model evidence, independent corroborating evidence and decision consequences justify this rating	Document the contribution of model evidence relative to other evidence and justify the assessment activities	Physician review or lack of autonomy alone does not establish medium risk
Low Risk	A COU assessed as low risk from model influence and decision consequence	Assessment activities proportionate to the documented COU and risk	Uses excluded from the draft's scope should be distinguished from in-scope low-risk uses

These are conditional illustrations of the proposed framework, not fixed FDA classifications by task. Independent verification reduced model influence in FDA's manufacturing example; high decision consequence still yielded medium model risk. Credibility activities should be tailored to model risk and COU (FDA draft guidance, sections II and IV.A.3–4).

</details>

Development and Validation of AI Models in Drug Development

Once the question, COU, and risk are defined, sponsors advance to **validating the AI model** for its intended use (Steps 4–5). This involves rigorous testing of model inputs, algorithms, and outputs, and collecting evidence that the model performs reliably within its COU. Important domains of validation include:

Data Quality and Fitness for Use

High-quality training and test data are the foundation of credible AI. The guidance and experts stress that model results are only as good as the data they're trained on. Sponsors must demonstrate that their data are “**fit-for-use**” in the given context (www.jdsupra.com) (emergingaihub.com). This includes:

- **Relevant and Representative Datasets.** The training set should reflect the COU. For instance, a model for patient selection should be trained on data from a similar patient population (similar demographics, disease severity, etc.). If important subpopulations are missing, the model may not generalize. Sponsors should describe data sources (e.g. clinical trial registries, electronic health records, manufacturing sensors) and ensure they cover the scope of use.
- **Data Preprocessing and Cleaning.** Any cleaning steps (handling missing values, standardizing measurements, removing outliers) should be documented. Explain how and why data were filtered or transformed. For example, if the model uses lab values, describe normalization methods.
- **Bias and Variance Checks.** Because historical data may contain biases, sponsors should check and document any potential biases in the training data. For example, if an ML model classifies patient risk, ensure that the dataset did not underrepresent a particular demographic. The Emerging AI Hub advises including bias and fairness assessment in the plan (emergingaihub.com). Practical steps can include subgroup analysis (performance by age/gender/race) or techniques to de-bias data.
- **Data Integrity and Traceability.** The draft recommends describing how development data are collected, processed, annotated, stored, controlled and used for training and tuning (FDA draft guidance, section IV.A.4.a.ii).

In regulatory submissions, sponsors will likely need to **describe the data** used to build the model. This could include data dictionaries, summaries of key variables, and sources. They should explain why these data are sufficient: e.g., “our training data encompassed 10,000 patients from four international trials, closely matching

the Phase 3 study population”. Any limitations or assumptions (such as ignoring rare events) should also be acknowledged, as these affect credibility.

Model Development and Architecture

Sponsors should detail how the AI model was built (Step 4). This covers the selection of algorithms, hyperparameters, model complexity, and software environment. While proprietary methods may remain confidential, the FDA guidance implies that *high-risk* models may require disclosure of the architecture. For example, the guidance notes disclosing model architecture and training methods “for high-risk AI models” (www.jdsupra.com). Even if exact code is not submitted, sponsors should document:

- **Algorithmic Approach.** What type of model is used (e.g. random forest, neural network, gradient boosting)? Why was this approach chosen for the COU?
- **Feature Selection.** Which input variables feed into the model, and how were they selected? (In images, this might be pixel data or features; in tabular data, which lab tests, vitals, etc.).
- **Hyperparameters and Training Process.** General description of how the model was trained: scores used, regularization to avoid overfitting, random seeds for reproducibility, number of training epochs, etc. While not all details need public disclosure, internal documentation should record the training process.
- **Data Partitioning.** How was data split into training/validation/test sets? Was cross-validation used? For example, in trial modeling: “We held out 20% of data from trial A as a validation set, and tested the final model on independent Phase 2 data from trial B.” This helps FDA see that the model was not just tuned to one dataset.

If the AI is a combination or hybrid model (e.g. combining an empirical physical model with an ML component), sponsors should describe this hybrid structure. Ark’s explanatory article uses an illustrative monoclonal-antibody process-development scenario ([Ark Biotech](#)). In such cases, clarifying how the pieces interact is important, as the joint credibility depends on both parts.

Foley LLP notes that extensive transparency is needed: “comprehensive details regarding the AI model’s architecture, training methodologies, validation processes, and performance metrics” may have to be submitted (www.jdsupra.com) for high-risk uses. In practice, sponsors should prepare technical documentation (like a model development report) that includes block diagrams or flowcharts of the model, even if not submitted (copies may be requested). Public summaries can be provided with as much detail as possible without revealing proprietary secrets.

Performance Validation (Steps 4–5)

Step 5 of the proposed framework involves **executing the credibility assessment plan** ([FDA draft guidance](#)). This means rigorously testing the model and quantifying its performance. The FDA expects a thorough demonstration that the model works as intended in the COU. Key elements include:

- **Independent Test Data.** The draft recommends test data that are independent of development data and are not shown to the algorithm during training ([FDA draft guidance, section IV.A.4.b](#)).

- **Performance Metrics.** Choose metrics appropriate to the decision task. Common metrics include accuracy, sensitivity/specificity, positive predictive value, AUC-ROC, mean squared error, etc. For classification tasks, a confusion matrix is useful. For probabilistic outputs, calibration plots or Brier scores can show reliability. Whatever metrics are used, sponsors should justify why they align with the regulatory goal. In a safety context, one might prioritize sensitivity (catching all true positives), whereas in a resource-constrained context, one might balance specificity.
- **Statistical Uncertainty.** It is crucial to quantify uncertainty in performance estimates. Provide confidence intervals or statistical significance for metrics. For example, if an accuracy is reported as 92%, also state the 95% CI (e.g. $\pm 2\%$). If possible, use bootstrapping or other resampling to estimate variability. This shows the range of possible performance and is part of “credibility evidence” ([regulations.justia.com](https://www.regulations.justia.com)).
- **Bias and Fairness Analysis.** Demonstrate that performance is consistent across relevant subgroups. If the model will be applied to all adults, show that it works similarly for men and women, or across age groups, etc. Any systematic discrepancies should be investigated. The guidance underscores bias checks as part of risk management ([emergingaihub.com](https://www.emergingaihub.com)). Note: absence of bias is not required for credibility, but sponsors must show they analyzed it.
- **Corner Cases and Robustness.** Test how the model performs in edge cases. For example, intentionally input unusual but plausible data to see if the model fails. Document any failure modes. Such robustness tests (sometimes called “sensitivity analyses”) can uncover weaknesses. Since AI models can sometimes give unpredictable results outside their training range, it’s prudent to test inputs marginally outside the expected distribution.
- **Comparison to Alternatives.** When relevant, compare the AI model’s performance to traditional methods. If a conventional algorithm or human decision is the gold standard, show that the AI is at least as good (or better). This contextualizes the benefit of AI. For example, if AI predicts clinical outcomes, compare against logistic regression or physician opinion.
- **Reproducibility Checks.** Run the model multiple times (if stochastic) to show results are stable. Save random seeds or apply techniques (e.g. fixed seeds or deterministic training pipelines) that ensure reproducibility of results, especially if code is audited.

All results should be meticulously **documented** (Step 5). Assess electronic records and signatures under the relevant predicate rules before assigning Part 11 controls; FDA recommends documenting applicability decisions ([FDA scope guidance](https://www.fda.gov/oc/ai/ai-scoping-guidance)).

Collectively, the validation evidence populates the credibility assessment report (Step 6). For example, Podichetty et al. suggest that high-risk AI submissions should include “model architecture, training logs, and data processing pipelines” in the documentation ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/36111111/)). In simpler terms, the FDA expects to see “(1) the model, (2) the data used to develop the model, (3) model training, and (4) model evaluation including test data and performance metrics” (www.jdsupra.com). These four components align with common regulatory sections (methods, data, validation plan, results) in an application, supplemented by AI-specific details.

Cross-Cutting Considerations: Uncertainty, Bias, and Explainability

Beyond basic validation, the FDA framework encourages sponsors to address certain cross-cutting issues that bear on credibility:

- **Uncertainty Quantification.** As noted, any single metric is an estimate, and AI models have inherent uncertainty. Sponsors should propagate and report uncertainty. This might include confidence intervals on predictions or probability outputs. For instance, if predicting a lab value, provide not just a point prediction but also a prediction interval. Techniques like Bayesian modeling or ensemble methods can yield measures of confidence. Transparency about uncertainty is important for FDA reviewers to judge reliability.
- **Algorithmic Bias.** ML models can inadvertently perpetuate biases present in training data. The FDA explicitly mentions assessing bias as part of credibility planning ([emergingaihub.com](https://www.emergingaihub.com)). For example, if an AI model for dermatology was trained mostly on light-skinned patients, its credibility would be suspect on darker skin. Sponsors should analyze performance by subgroup (race, gender, age, etc.) and either enrich the data or qualify the model's limits. If biases are found, discuss mitigation strategies (reweighting, separate models, disclaimers). Documenting that the model does not have clinically significant biases (or explaining any discovered bias) strengthens credibility.
- **Explainability and Transparency.** While not a formal requirement in drug applications, explainability aids trust. The draft guidance does not mandate full interpretability for all AI, but Foley points out that one innovation need is “model explainability” (www.foley.com). Providing model rationales (feature importance, attention maps) can help FDA reviewers and investigators accept the results. Sponsors should at least document how the model makes decisions at a high level. For instance, decision-tree models have natural interpretability; neural networks might use SHAP values or similar.
- **Regulatory Compliance (Part 11, GxP).** Assess Part 11 applicability record by record under the relevant predicate rules, including required retention or submission and reliance on electronic records. Apply the controls and signature requirements that govern those records ([FDA scope guidance](#)).

Step 4–7: Planning, Execution, Documentation, and Review

Following risk assessment, Steps 4–7 formalize how to demonstrate credibility:

- **Step 4: Develop Credibility Plan.** Sponsors create a structured plan addressing all aspects of evidence collection. It should cover data quality checks, modeling strategy, performance criteria, validation design, and bias analysis ([emergingaihub.com](https://www.emergingaihub.com)). Known as a “credibility assessment plan”, this document is akin to a validation protocol in clinical trials. It should specify acceptance criteria (e.g. “sensitivity $\geq$ 90% with 95% CI enclosed in trial standard”), data requirements, and responsibilities. The DLA Piper analysis emphasized that the entire framework’s purpose is to “establish trust in the AI model by collecting credibility evidence...for its particular COU” (www.dlapiper.com), and the plan lays out exactly how that evidence will be obtained.

- **Step 5: Execute the Plan.** Run the planned analyses and record results and deviations. Determine which electronic records and signatures are covered by Part 11 under the applicable predicate rules ([FDA scope guidance](#)).
- **Step 6: Document Results and Deviations.** The draft recommends documenting results and deviations in a credibility assessment report. Sponsors should discuss with FDA whether, when and where to submit the report ([FDA draft guidance, section IV.A.6](#)).
- **Step 7: Determine Adequacy for COU.** Under the proposed framework, Step 7 considers whether the documented results support the model's use for the COU ([FDA draft guidance, section IV.A.7](#)). Adequacy is not a binary pass/fail, but a judgment call considering the model's demonstrated performance in light of risk ([emergingaihub.com](#)). For example, a model with 95% accuracy might be adequate for a low-risk contingency plan but inadequate if missing the 5% leads to serious patient harm. The sponsor should explicitly state its conclusion: "Based on the evidence, the model is (or is not) sufficiently credible to support [the COU]." If it is not, additional work (e.g. more data, model revision) is needed before submission.

Section IV.B recommends **ongoing monitoring** of model performance, with oversight commensurate with model risk and COU ([FDA draft guidance, section IV.B](#)).

By completing these steps and assembling a thorough report, sponsors can demonstrate to FDA that they have "systematically built, documented, and evaluated" the model's credibility ([emergingaihub.com](#)). In sum, establishing credibility means **multiple layers of evidence**: strong data, robust performance testing, risk justification, and transparent reporting. The process mirrors that for any critical tool in drug development, with added attention to AI-specific issues.

Data Analysis and Evidence – Building Credibility

Beyond the structured steps, it is useful to examine the underlying **data analysis practices** that strengthen an AI model's credibility. Here we consider how to handle data, measure performance, and interpret results in depth.

Ensuring High-Quality Training Data

- **Data Source Evaluation:** Document the origin of every dataset used. Did it come from historical trials, real-world registries, manufacturing sensors, or other sources? Ensure these sources are reliable and relevant. For example, if training a model for a new diabetes drug trial, using EHR data from diabetic patients (rather than general population data) makes the model more credible.
- **Data Completeness and Labeling:** Check for missing or mislabeled data. For supervised models, labels must be accurate (e.g. true outcomes). If labels come from imperfect tests or human adjudicators, note their accuracy. Missingness should be handled systematically (e.g. imputation methods). The sponsor should report how many records were excluded for missing data or poor quality.

- **Preprocessing Procedures:** Any normalization, encoding (e.g. one-hot for categories), or filtering applied to the data must be described. For instance, continuous lab values might be log-transformed; imaging pixels might be standardized. Provide rationales (e.g. “log-transformation reduced skewness”). This assures FDA reviewers that the data fed into the AI were reasonably conditioned.
- **Feature Engineering:** If new features are derived (e.g. combining blood pressure and heart rate into a risk score), explain how. Does the model purely learn from raw inputs, or were domain transformations applied? Documenting feature engineering helps others understand model drivers.
- **Training/Test Splits:** Clearly specify how data were partitioned. A common best practice is 70% training, 30% testing, or k-fold cross-validation. Identify any external datasets if used for testing. For example, “Data from Trials A/B were used for training; Trial C data held out for final validation.” This prevents “data leakage” where test data inadvertently influence training.
- **Size and Diversity:** State the sample size and diversity of the training data. AI models often require large datasets to generalize. If data are limited, consider data augmentation (for images) or note that limitations exist. For credible submission, extreme paucity of data should be justified and perhaps mitigated by conservative performance claims.

Performance Evaluation

- **Metrics Selection:** Choose metrics that reflect the regulatory question. For categorical outcomes, measures like sensitivity (recall), specificity, and predictive values are typically reported. For regression tasks (e.g. predicting a concentration), use mean absolute error or R^2 . Complex outcomes (like survival) might require concordance indices or Kaplan-Meier calibration plots. Always tie metrics to clinical meaning.
- **Summary Statistics:** Provide the mean, median, standard deviation, etc., of the model’s outputs on test sets. For classification, provide a confusion matrix to show true positive/negative and false positive/negative counts. For probabilities, consider calibration plots. Clear tables of results allow others to gauge performance at a glance.
- **Visualizations:** Whenever possible, include figures such as ROC curves, precision-recall curves, or predicted-vs-actual scatter plots. These visuals communicate performance nuances and can highlight any systematic issues. For instance, an ROC curve can show trade-offs between sensitivity and specificity.
- **Uncertainty and Confidence:** As noted, give confidence intervals or p-values to accompany point estimates. FDA reviewers will appreciate knowing the statistical confidence of performance claims. If a subgroup analysis is done (e.g. model accuracy by age group), give separate intervals.
- **Benchmark Comparisons:** If relevant, compare the AI model against existing benchmarks or conventional methods. For example, show that a neural network outperforms logistic regression or a simpler rule-based algorithm. If AI is replacing human review, compare to inter-observer variability. Demonstrating superiority (or at least equivalence) to the status quo helps justify novelty and credibility.

- **Ad Hoc and Sensitivity Analyses:** Document any additional analyses done to test model stability. For example, “We retrained the model 100 times with different random seeds; accuracy varied by $\pm 1\%$. We also tested the model by adding 5% noise to inputs to simulate measurement error; performance only dropped from 92% to 89%.” These sensitivity checks reveal robustness.

Dealing with Bias and Fairness

Bias in AI is a major concern. Sponsors should proactively address it:

- **Identify Protected Attributes:** Determine which demographic or medical subgroups are relevant (race, sex, age, disease subtype, etc.).
- **Stratified Performance:** Evaluate and report model performance for each subgroup. For example, present accuracy for male vs. female, or for each ethnic group. If a model fares significantly worse in any subgroup, investigate why.
- **Balance or Reweight Data:** If a disparity is found due to underrepresentation (e.g. only 10% of training data are from group X), consider augmenting data or using techniques like SMOTE. Discuss any steps taken to balance.
- **Fairness Metrics:** Consider metrics like equalized odds or demographic parity if applicable. While not yet standardized in drug regulation, showing awareness of fairness principles (for example, by ensuring false positive rates are similar across groups) strengthens credibility.
- **Transparency of Limitations:** If no mitigation is possible (e.g. simply no data exist for a subgroup), clearly note it. FDA reviewers need to know the limits of generalization. In some cases, the sponsor may restrict the COU (e.g. not claiming applicability to under-tested groups).

Documentation of Analysis

All analysis steps should be **logged and reproducible**:

- **Code and Environment:** Maintain the exact code (scripts, notebooks) used for training and testing. Although submitting proprietary code may not be required, having a repository with version control (e.g. Git) provides traceability. Record software versions of libraries and tools, so that an exact environment can be recreated (or simulated by FDA if needed).
- **Randomness Control:** If the model training has stochastic elements (random weight initialization, random data splits), ensure results are reproducible by fixing seeds or documenting the random state. This is part of *practical reproducibility*.
- **Record of Changes (Change Log):** Document every change made to the model after initial training. For example, if hyperparameters were tweaked, note the change and its effect. This log of modifications will feed into the credibility report as the project evolves.

- **Quality Management Integration:** Document which AI records fall within Part 11 and which signatures the predicate rules require. Select traceability controls using the applicable rules and a documented risk assessment ([FDA scope guidance](#)).

Real-World Case Studies

The cases below distinguish an illustrative industry scenario from documented regulatory analyses.

Case 1: AI in Bioprocess Optimization (Ark Biotech)

Scenario: Ark Biotech's explanatory article presents an illustrative continuous monoclonal-antibody process-development scenario focused on optimizing perfusion rate ([Ark Biotech](#)). Read it as a hypothetical teaching example, without attributing completed validation studies or regulatory filings to Ark.

Case 2: AI for Patient Selection (COVID-19 Trial)

Scenario and COU: FDA's November 8, 2022 EUA covered Kineret (anakinra) for COVID-19 in hospitalized adults with positive direct SARS-CoV-2 tests and pneumonia requiring supplemental oxygen, at risk of severe respiratory failure and likely to have elevated suPAR. SCORE2 supported identification of that population ([FDA review](#)).

Data and Validation: During EUA review, CDER developed the scoring rule from SAVEMORE data and externally validated it with SAVE data to identify likely suPAR ≥ 6 ng/mL because a commercial test was unavailable in the U.S.

Limitations: Development was exploratory and sensitivity was low; patients who might benefit could be missed. Benefit among score-negative patients remained uncertain. This was a retrospective regulatory analysis, not a prospectively tested adaptive trial design ([FDA research account](#)).

Case 3: AI-Enabled Image Analysis for Endpoints (NASH Trial)

Scenario: PathAI's AIM-NASH assists pathologists scoring liver biopsies in MASH trials. EMA announced CHMP qualification on March 20, 2025 ([EMA announcement](#)).

Qualified COU: The opinion covers an aid to a single central pathologist for enrolment in phase 2 and 3 MASH trials and histology-based outcome evaluation. The pathologist actively reviews and accepts or rejects the scores; slides must be scanned with the Aperio AT2 ([EMA qualification opinion](#)).

FDA Status: FDA announced AIM-NASH qualification on December 8, 2025, for assessing MASH disease activity in clinical trials. Pathologists remain responsible for final interpretation ([FDA qualification announcement](#)).

Documentation and Regulatory Submission

The draft recommends documenting the results of the credibility assessment plan and deviations from it ([FDA draft guidance, section IV.A.6](#)). In a submission (e.g. IND, NDA, BLA filings), sponsors should include an AI addendum or sections in relevant places (clinical protocols, CMC sections, etc.) describing the AI model. Key documentation items include:

- **Model Description:** A clear narrative of what the model is and does (COU, question it answers). This may go in the clinical or quality sections depending on use. Provide a summary of its design and purpose.
- **Inputs and Outputs:** Define all inputs (features, sensors, etc.) and outputs. For instance, if the AI works on MRI scans, describe the imaging protocol and preprocessing. If it ingests blood test results, list them. Also specify how outputs are interpreted (e.g. risk score thresholds).
- **Training and Validation Data:** Describe the datasets used (size, source, features) for training and validation. If using proprietary datasets (e.g. from prior trials), explain their provenance and relevance. Indicate which data were internal vs. any external or public sources.
- **Performance Metrics and Results:** Provide tables/figures of model performance on test sets, including metrics (accuracy, sensitivity, etc.) and statistical confidence. Include subgroup analyses (e.g. by relevant patient characteristics). This helps demonstrate reliability.
- **Risk Analysis:** Summarize the risk assessment (Step 3). This can be a simple table or paragraph explaining the chosen influence/consequence levels and justification. It should show that the sponsor has considered the implications of model failure.
- **Validation Procedures:** Outline the specific validation studies performed. E.g. "The model was evaluated on a blinded external dataset of 200 patients, achieving X% accuracy. We also retrospectively applied the model to historical trial participants and confirmed that treatment effect was predicted correctly." Even if not full clinical trial validity, showing the model reproduces known findings boosts confidence.
- **Change Control Plan:** For manufacturing uses, evaluate model changes through the pharmaceutical quality system's change management process. Reassess credibility and repeat relevant assessment activities as needed; determine and meet applicable reporting requirements based on the change and its impact ([FDA draft guidance, section IV.B](#)).
- **Software Validation:** If submitting software (e.g. algorithm as part of a device or analytical package), include evidence of software testing and qualification. Otherwise, at least assert that the computational infrastructure is validated for its intended use.
- **Transparency vs. IP:** Disclosure to FDA is distinct from public disclosure. Under [21 CFR 20.61\(c\)](#), submitted information qualifying as a trade secret or confidential commercial or financial information is protected from public disclosure. Obtain case-specific legal advice on patent and trade-secret strategy.

- **Integration with Quality Systems:** Document how the AI development project fits into the company's quality management. This might include references to GLP/GMP standards, computer system validation records, or good machine learning practice (GMLP) guidelines. Showing that the AI code and data are managed under QA controls earns trust.

Overall, the goal is for a regulatory reviewer to **follow the train of logic**: from COU to risk to validation to conclusion. The documentation package should include a written trace of all steps and decisions (much like a clinical study report). Reviewers should not have to guess how the model was tested or where results came from. As Emerging AI Hub emphasizes, transparency is key: "The documentation package should be structured so that an FDA reviewer can independently evaluate the model's credibility without needing to rerun any analyses" (emergingaihub.com).

Implications, Challenges, and Future Directions

The FDA's AI credibility guidance represents a major milestone. It signals that AI/ML is maturing from experimental to regulated territory in drug development. However, it also raises significant implications and challenges:

- **Innovation Opportunities.** The draft offers recommendations rather than establishing legally enforceable responsibilities ([FDA draft guidance](#)). Foley identifies opportunities such as explainable AI techniques, automated bias-detection tools, and systems for model monitoring and reporting (www.foley.com). For example, "automatic systems to generate reports of model development and evaluation" are specifically cited, since sponsors will need to compile lots of documentation (www.foley.com). Companies that develop tools for AI transparency or life-cycle management (e.g. drift detectors, automated audit logs) could become valuable partners. Cryptographic methods (federated learning, homomorphic encryption) might help protect data IP under these disclosure rules.
- **Intellectual Property Tensions.** Supplying information for FDA review does not itself mean publishing proprietary methods. [21 CFR 20.61\(c\)](#) protects submitted information that qualifies as a trade secret or confidential commercial or financial information. Patent-versus-trade-secret decisions call for case-specific legal advice.
- **Regulatory Harmonization.** The FDA is not alone in addressing AI. The European Medicines Agency (EMA) published a reflection paper in September 2024 on AI in the entire medicine lifecycle (www.ema.europa.eu), emphasizing a risk-based, human-centric approach. EmergingAIHub notes that in January 2026 FDA and EMA released joint principles for "good AI practice" in drug development (emergingaihub.com). There is also work from standards bodies (e.g. ISPE's July 2025 GAMP guide on AI in GxP). While details differ, a convergence is apparent: risk-based validation, transparency, and post-market monitoring are common themes. Sponsors should breathe a sigh of relief that global regulatory environments seem to be coalescing rather than diverging. Preparing for FDA's framework will likely put a company in good position for EMA or other regulatory bodies.

- **Operational Burden.** The compliance cost will be non-trivial. Building a comprehensive justification and documentation for an AI model rivals that for a drug formulation. Smaller biotech firms may find the burden heavy; they may need to partner with AI service providers who specialize in regulatory-grade ML. It will also require cross-functional teams: data scientists, statisticians, clinicians, and regulatory affairs experts must collaborate. Training programs on “regulatory ML” may emerge.
- **Data Considerations.** The emphasis on data quality highlights an ongoing need for better data infrastructure. Poor data can sink AI credibility. Regulators may therefore push for centralized data standards (like common formats for electronic health records) to ease future AI validation. Companies may invest more in data curation and bias audits, even outside regulated contexts.
- **Future FDA Actions.** The original notice requested comments by April 7, 2025, to ensure consideration before work on finalization ([Federal Register notice](#)). FDA still lists the document as draft guidance, not for implementation, and permits comments at any time ([FDA guidance page](#)).
- **Case Study Evidence Base.** As more AI applications reach submissions, FDA reviewers will accumulate experience. In its January 6, 2025 announcement, FDA reported experience with more than 500 drug and biological product submissions with AI components since 2016 ([FDA announcement](#)). This means FDA staff are learning what works and what problems arise. The feedback cycle (comments on the draft) will also shape future policies. Sponsors should monitor FDA’s published decisions (e.g. summary reviews) to glean how the credibility framework is being applied in reality.
- **Ethical and Societal Implications.** Beyond technicalities, credible AI in healthcare raises questions of trust and consent. Patients and clinicians may be wary of “black box” models. The emphasis on documentation and explainability can help build public trust. Moreover, transparent validation reports could end up as part of regulatory records accessible to healthcare professionals, not just regulators. In that sense, the guidance nudges the industry toward a more open model of algorithmic accountability.

Conclusion

The FDA document remains draft guidance, is not for implementation, and contains nonbinding recommendations ([FDA guidance page](#)). Existing statutory and regulatory requirements still apply; the draft does not create new legally enforceable responsibilities. The seven-step framework provides a clear (if demanding) roadmap: define a specific question and context, analyze risks, and then validate the model to a degree commensurate with those risks ([regulations.justia.com](#)) ([emergingaihub.com](#)). Throughout, the focus is on *credible evidence*: high-quality data, robust testing, and meticulous documentation.

This report has unpacked the draft’s proposed recommendations in detail. We traced the historical impetus (ever-increasing AI use in submissions ([www.fda.gov](#)) ([emergingaihub.com](#))), explained each step of the new guidance, and highlighted key data and validation practices. Multiple viewpoints (legal, technical, academic) and real examples demonstrate how to apply these principles in practice. Importantly, we saw that credibility is holistic: it is built not only on one metric but on a chain of rigor – from how you frame the question to how you monitor model performance over its life cycle ([FDA draft guidance, section IV.B](#)).

In summary, evaluating and documenting an AI model for drug development means treating it like any critical scientific instrument. Sponsors must plan from the outset: choosing appropriate data, selecting performance goals, and recording every step. They must be ready to explain the AI to regulators comprehensively, often in ways they never had to for traditional statistical models. As FDA's Commissioner Califf noted, the goal is to foster innovation while upholding safety and efficacy standards (www.fda.gov). By following the FDA's framework, developers can embrace AI's promise – using it to accelerate drug development and improve care – with the safeguards needed for public trust.

Sources: This report draws on FDA statements and draft guidance (www.fda.gov) (regulations.justia.com), peer-reviewed analyses (academic.oup.com) (pmc.ncbi.nlm.nih.gov), legal and industry commentary (www.jdsupra.com) (www.dlapiper.com), and relevant case studies (www.ark-biotech.com) (pmc.ncbi.nlm.nih.gov). All claims above are supported by cited evidence. Any quotations from FDA or expert sources are explicitly referenced with hyperlinks to the source.

Source links

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

1. <https://regulations.justia.com/regulations/fedreg/2025/01/07/2024-31542.html>
2. <https://emergingaihub.com/ai-compliance/fda-ai-credibility-framework-guide>
3. <https://www.fda.gov/media/184830/download>
4. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12675821>
5. <https://www.foley.com/de/insights/publications/2025/01/ai-drug-development-fda-releases-draft-guidance>
6. <https://www.ark-biotech.com/insights/making-pharma-ai-ready-applying-the-fdas-draft-guidance>
7. <https://www.fda.gov/media/185783/download>
8. https://www.ema.europa.eu/en/documents/other/qualification-opinion-artificial-intelligence-based-measurement-non-alcoholic-steatohepatitis-histology-liver-biopsies-determine-disease-activity-nash-mash-clinical-trials_en.pdf
9. <https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle-scientific-guideline>
10. <https://www.jdsupra.com/legalnews/ai-drug-development-fda-releases-draft-3721117>
11. <https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions>
12. <https://academic.oup.com/jlb/article-abstract/doi/10.1093/jlb/laf028/8316994>
13. <https://www.dlapiper.com/en-it/insights/publications/2025/01/fda-releases-draft-guidance-on-use-of-ai>
14. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>
15. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application>
16. <https://www.fda.gov/drugs/spotlight-cder-science/using-machine-learning-identify-suitable-patient-population-anakinra-treatment-covid-19-under>
17. <https://www.ema.europa.eu/en/news/ema-qualifies-first-artificial-intelligence-tool-diagnose-inflammatory-liver-disease-mash-biopsy-samples>

18. <https://www.fda.gov/drugs/drug-safety-and-availability/fda-qualifies-first-ai-drug-development-tool-will-be-used-mash-clinical-trials>
19. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-20/subpart-D/section-20.61>
20. <https://www.federalregister.gov/documents/2025/01/07/2024-31542/considerations-for-the-use-of-artificial-intelligence-to-support-regulatory-decision-making-for-drug>

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.