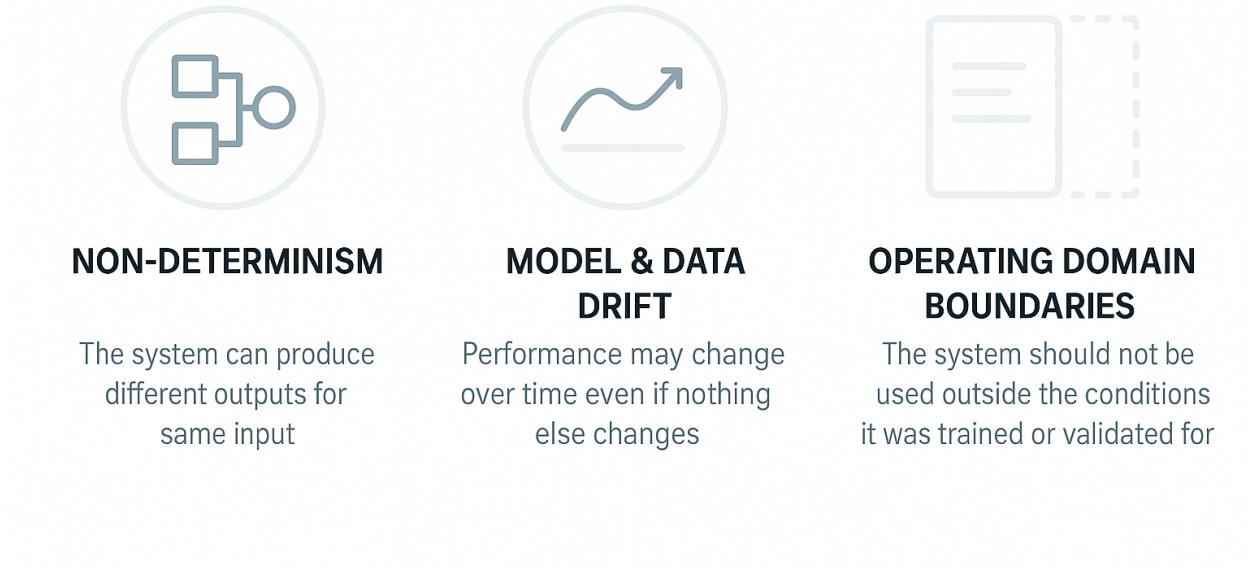
Validating AI Systems in GMP:

A Beginner's Guide for CSV Professionals



How to Use This Guide: *I've structured the guide into seven parts, each focusing on a key aspect of AI system validation in GMP. You'll find simple explanations of AI terms, examples (e.g., real use cases in labs and manufacturing), and tips drawn from current regulations and industry best practices. This is built upon regulatory documents (FDA, EMA, PIC/S, ISPE) with inline references (e.g., "(2025, FDA Draft Guidance)") so you know where recommendations are coming from. Full reference details are provided at the end with URLs for further reading. Also, abbreviations or technical terms are defined in the* ***Glossary*** *in Part 7 for reference..*

Above all, this guide is meant to be practical. It's written for CSV professionals. By the end, you should feel confident explaining AI concepts to your team, planning an AI validation project, and ensuring compliance with the latest GMP expectations. Let's get started!

1. Introduction

Welcome to validating Artificial Intelligence (AI) systems in Good Manufacturing Practice (GMP) environments. In this guide, I'll walk you through the very basics of AI to the specific compliant practices needed to satisfy GMP regulators. We'll build from foundational concepts up to applied strategies that you can immediately use in your validation projects.

Table of Contents

Table of Contents Summary

1. Introduction to the guide and intent.

2. Foundations of AI for CSV Professionals – Explains what AI and Machine Learning are in simple terms, how they differ from traditional software, and why those differences matter for validation.

3. Regulatory Landscape Overview (FDA, EMA, ISPE, PIC/S) – Summarizes key regulations and guidances from the FDA, EMA, ISPE's GAMP guide, and PIC/S that specifically address AI in GMP. We'll highlight what regulators expect and the principles you must align with (e.g., risk management, transparency, and human oversight). Also refer to Appendix H.

4. AI Lifecycle Explained Step by Step – Breaks down the life cycle of an AI system, from concept and data gathering through model development, validation, deployment, and ongoing monitoring. We'll map these steps to familiar CSV life cycle phases and discuss deliverables at each stage.

5. Risk Management for AI in GMP – Focuses on how to assess and control risks unique to AI systems. We'll use a Quality Risk Management approach (ICH Q9 principles) to identify potential failures (like biased data or model drift) and implement mitigations (like human-in-the-loop review and drift monitoring).

6. Validation Framework for Decision Support Tools and Lab Automation – Provides a practical framework to validate AI-powered decision support systems (where AI assists humans in making decisions) and AI in lab automation (such as intelligent lab instruments or robotics). Includes use case examples, validation testing tips, and how to document evidence for audits.

7. Templates, Glossary, and References – Offers suggestions for templates or documentation (like an "AI Validation Plan" outline or a "Model Test Results" template), a glossary defining all key terms and abbreviations used in this guide, and complete reference listings for all regulations and sources cited.

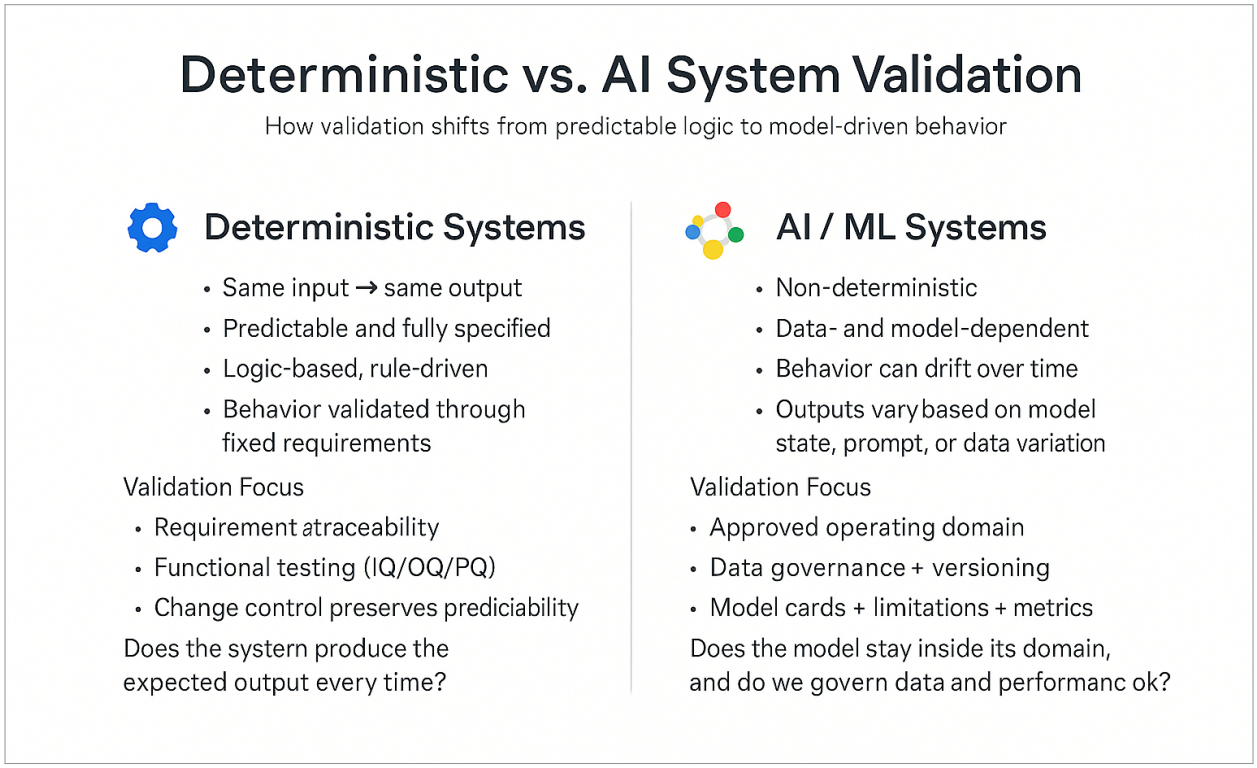
Let's dive into the foundations of AI to build a common understanding before we tackle compliance and validation strategies.

2. Foundations of AI for CSV Professionals

What Do "AI" and "ML" Mean? Terms like *Artificial Intelligence (AI)* and *Machine Learning (ML)* are thrown around but how are they different from regular software? In simple terms, **AI** is a broad concept referring to machines performing tasks that would normally require human intelligence (like decision-making or pattern recognition). **Machine Learning** is a subset of AI – it's a way to create AI systems by feeding them data and letting them learn patterns on their own, rather than programming every rule explicitly. In other words:

- *Traditional software* is programmed with explicit instructions: if X, do Y. Every possible scenario is anticipated by developers (think of an Excel formula or a piece of validation script – it will do exactly what you coded it to do, nothing more).
- *Machine Learning software* is *trained* rather than explicitly programmed. Developers create an algorithm that can improve its performance by learning from data. For example, instead of coding specific rules to detect a defective tablet image, we feed thousands of images to a machine learning model and it "figures out" the distinguishing features of a defective tablet vs. a good one.

Why This Matters for Validation: In traditional CSV, we start with fixed user requirements and we verify the system meets those requirements through testing. With AI/ML, the system's behavior isn't entirely fixed by code; it's also determined by training data and the learning process. This introduces two big differences for us as validators:



1. **Data is King:** The quality and representativeness of the data used to train and test an AI model directly affect the system's performance. If the training data has gaps or biases, the model's outputs will reflect that. As a validator, I now must scrutinize data with the same rigor as software requirements. We need to ask: *Were the training, validation, and test datasets properly curated and representative of real-world conditions?* (we'll cover how to do this in the lifecycle and risk sections.)

2. **The System Can "Learn":** Some AI models can adapt over time (learn continuously) or be re-trained with new data. This is very different from a static system that only changes when a developer releases an update. An AI system's performance might improve or drift without a formal software change. For GMP, this is a potential nightmare if not controlled – imagine an algorithm gradually drifting until it mis-classifies a quality result. In validation, we have to establish procedures to detect and manage any such changes (e.g., periodic re-validation or monitoring for drift). In fact, regulators currently *discourage* using continuously self-learning models in critical GMP processes because of the difficulty in controlling them. For instance, the upcoming EU GMP Annex 22 explicitly states that **"dynamic models which continuously and automatically learn… should not be used in critical GMP applications"** (2025, EMA Draft Annex 22). Instead, any learning should be done in a controlled retraining process, not on-the-fly in production.

Key AI Concepts Explained in Plain English

Before going further, let's clarify a few terms you'll encounter:

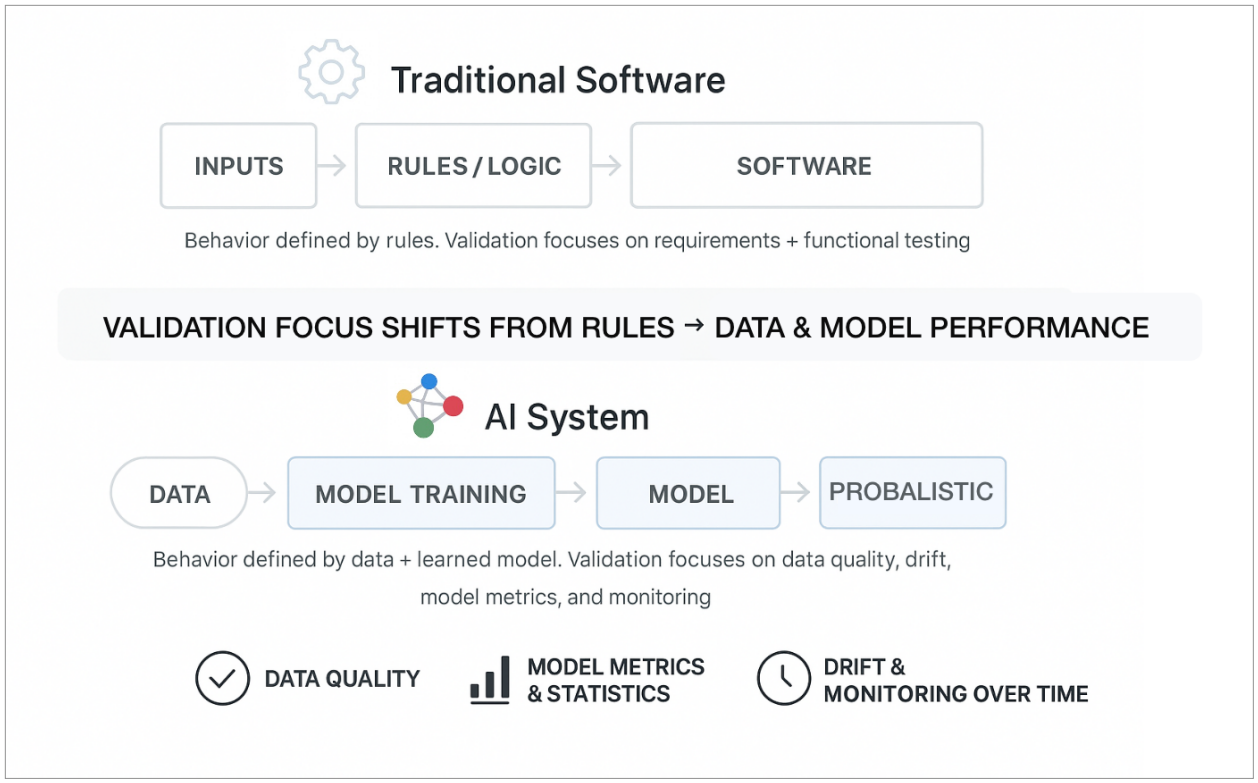
- **Algorithm vs. Model:** An *algorithm* is the procedure or recipe (often math or code) for learning from data. A *model* is the result of that learning – essentially the trained software artifact that can make predictions or decisions. For example, a developer chooses a machine learning algorithm (say, a neural network). After training it on data, the finished neural network (with all its weights set) is the model. In validation, we often care about both: Was the algorithm appropriate and validated? And, is the trained model performing as expected?

- **Training vs. Testing:** *Training data* is used to teach the model (adjusting its internal parameters to fit patterns). *Testing data* (or validation data) is separate data used to evaluate how well the trained model performs on new, unseen examples. A golden rule in ML is to keep test data independent – never let the model train on what you plan to test it on, otherwise the evaluation is biased. We CSV folks can relate this to not "teaching to the test"; it's like reserving a set of requirements to challenge the system that weren't explicitly coded for.

- **Overfitting:** This is when a model memorizes the training data too closely and fails to generalize to new data. It's like a student who memorized practice exam answers but can't handle a slightly different real exam question. During validation of AI, we check

for overfitting by seeing if the model's accuracy on training data is much higher than on test data – a red flag that it might not perform well in production.

- **Bias:** In AI, bias doesn't just mean prejudice; it refers to systematic errors in the model's output due to skewed data or design. For instance, if all your training data for an assay analyzer came from one lab instrument, the AI might be biased to that instrument's characteristics and perform poorly on another instrument. Regulators emphasize preventing and detecting bias – *"active measures must be taken to minimise the integration of bias into AI/ML applications and promote reliable and trustworthy AI"* (2024, EMA Reflection Paper). Refer to Appendix E for more on bias).
- **Explainability:** This is the ability to understand and explain how an AI model makes decisions. Traditional software logic can usually be traced step-by-step, but complex AI models (like deep neural networks) are often "black boxes." In GMP, we don't necessarily need to know the algorithm's inner math if we can prove it works reliably. However, having some level of explainability builds trust – both for regulators and for users. For example, if an AI flags a batch as anomalous, can it also highlight which parameters were most influential? Annex 22 specifically mentions *explainability* as a consideration, meaning you should choose or design models that are interpretable enough for validation purposes (2025, EMA Draft Annex 22).



Connecting to What You Already Know

If all these terms feel abstract, it might help to draw parallels with the familiar validation world:

- Writing an **algorithm** in AI is akin to writing functional specifications for a traditional system. It defines the approach but not the exact outcomes.
- **Training** an AI model is a bit like configuring a system or populating a database – you're providing the content that the system's behavior will depend on.
- **Testing** an AI model with a reserved dataset is analogous to performing qualification tests (IQ/OQ/PQ) on a system against predefined acceptance criteria.
- Controlling **model updates** is just another form of change control. If a model is retrained (say, to improve accuracy or adapt to new product data), treat it like a new software version: document the changes, assess the impact, run a regression test (i.e., test on new independent data), and get approval before deploying.

The core message in this section: **AI systems are still software systems, but with extra layers (data and learning) that we have to manage.** We don't throw out our CSV rulebook – we extend it. Throughout the rest of this guide, keep this foundation in mind. We'll now move on to the regulatory landscape to understand what our compliance targets are, and then we'll dive into how to actually carry out validation for AI step by step.

3. Regulatory Expectations for AI in GMP, Clinical, and PV

Purpose of This Chapter

The goal here is to give you a clear and practical understanding of what regulators expect when we deploy AI systems in GMP, clinical, and PV environments. I focus on the requirements that support real world validation work: human oversight, explainability, traceability, transparency, and continuous monitoring. These concepts show up consistently across FDA, EMA, PIC S, ISPE, and even the EU AI Act.

By the end of this chapter, you will know how to interpret the regulatory language and apply it when writing your validation plan, URS, risk assessments, and monitoring procedures.

Why AI Requires Its Own Regulatory Lens

Traditional software behaves the same way every time. AI models do not. They are statistical systems that generalize patterns and sometimes behave unpredictably, especially when exposed to new data. Regulators recognize this and expect additional controls that help us demonstrate credibility, maintain oversight, and prevent unexpected behavior.

In simple terms:
AI needs additional controls because its behavior can shift, and the reasons for those shifts are not always obvious.

Regulators focus on four themes:

1. Human oversight
2. Explainability
3. Scientific and analytical validity
4. Ongoing monitoring and drift detection

Everything else flows from these ideas.

Human Oversight (Human in the Loop)

Across all major regulators, the strongest and most consistent requirement is human oversight. This is usually referred to as **Human in the Loop (HITL).** HITL is the expectation that a human must be able to review, verify, and override any AI output before it influences a GMP relevant decision.

What regulators actually say

FDA: requires "appropriate human oversight" for AI used in regulatory decision making (2025, FDA Draft Guidance).

EMA: calls for a "human centric approach" and states that users must be able to override AI assisted outputs, especially in higher impact areas (2024, EMA Reflection Paper).

PIC S Annex 22: requires "oversight mechanisms" for high impact uses and states that AI outputs must not bypass human review in quality decisions (2025, PIC S Annex 22).

ISPE GAMP 5: identifies HITL as a primary control for nondeterministic and adaptive AI models (2023, ISPE GAMP Update).

EU AI Act: requires "effective human oversight" for high risk AI systems that impact safety or regulated decisions.

Why HITL matters in validation

HITL increases detectability and reduces the risk that an incorrect AI output influences a decision. It also ensures the final responsibility remains with the qualified person, not the model.

How it translates to validation deliverables

- URS: "The system shall require a qualified user to confirm or override AI suggested outputs before they influence any GMP decision."
- Test scripts: challenge tests where the model is wrong and the human catches it.
- SOPs: instructions for approving, rejecting, or overriding AI suggestions.
- Monitoring: capture override rate as a performance indicator.

Explainability and Transparency

Explainability means users must understand why the model produced a given output or at least understand the factors that influenced the prediction. Regulators do not require full mathematical transparency, but they expect explainability that is appropriate for the risk level.

What regulators say

FDA: expects "interpretability appropriate to the intended decision" (2025, FDA Draft Guidance).

EMA: states that explainability must be "proportional to risk" and that AI supported activities must provide enough transparency for the user to trust and verify the output (2024, EMA Reflection Paper).

PIC S: states that AI assisted decisions "must be explainable to personnel and inspectors" (2025, PIC S Annex 22).

ISPE GAMP 5: encourages using interpretable models or providing supporting explanations for higher risk uses.

Why explainability matters

Explainability gives humans the information they need to determine whether the model is behaving correctly. It also helps inspectors understand how a model fits within the quality system.

How it translates to validation deliverables

- URS: "The system shall provide supporting evidence or key factors that influenced the AI output."
- Validation: confirm that explanations align with domain knowledge.
- Monitoring: track cases where users report unclear or unhelpful explanations.

Credibility: Scientific, Analytical, and Process Validity

Regulators frame AI trustworthiness in terms of **credibility**. This means the model must be scientifically sound, statistically validated, and proven to work correctly in the real process it supports.

Scientific validity

The modeling approach must match the scientific or engineering problem. Regulators want to see that the model architecture and inputs make sense.

Analytical validity

This is the performance testing: accuracy, precision, recall, false negatives, robustness, calibration curves, stress testing, and edge cases.

Process validity

The model must work correctly in the actual business process. This is where pilot studies, user feedback, and SME confirmation come in.

Regulatory citations

- FDA requires evidence that AI systems are "valid, reliable, and appropriate for their context of use" (2025, FDA Draft Guidance).
- EMA calls for "scientific validity and reliability" for any AI system used across the medicinal product lifecycle (2024, EMA Reflection Paper).

- PIC S requires evidence of trustworthiness in the approved operating domain (2025, PIC S Annex 22). Refer to appendix D.

Drift and Continuous Monitoring

AI systems can degrade over time because data changes. Regulators expect continuous monitoring to detect drift early (appendix F).

Regulatory expectations

- FDA expects ongoing performance checks aligned to the model's context of use (2025, FDA Draft Guidance).
- EMA requires lifecycle based monitoring to detect degradation (2024, EMA Reflection Paper).
- PIC S Annex 22 requires monitoring for "unexpected behavior or performance degradation."
- ISPE GAMP 5 states that AI must be monitored after deployment to maintain its validated state.

How this shows up in validation

- define drift thresholds
- define monitoring frequency
- define triggers for retraining or CAPA
- log inputs, outputs, and overrides
- maintain version control for retrained models

Operating Domain and Context of Use

Regulators expect a clear statement describing where the model is allowed to operate and what tasks it supports. Validation is performed against that specific domain and no broader. COU is mentioned explicitly in FDA guidance and implicitly in EMA and PIC S documents.

Validation implication

If COU changes, validation must change with it.

What Inspectors Look for in AI Systems - Inspectors evaluate AI through a lens that is familiar to them**:**

- Is there human oversight
- Can users explain the output
- Is the model validated for its context
- Is the data appropriate
- Is performance monitored
- Are limitations documented
- Are version changes controlled
- Is the model traceable and challengeable

If these elements are in place, the system is usually accepted. Refer to Appendix H for more.

4. AI Lifecycle Explained Step by Step

When I first approached validating an AI system, I realized I needed to understand its *lifecycle*: not just the software development lifecycle (SDLC) we know, but the **model development lifecycle** that data scientists follow. What I discovered is that an AI system's lifecycle can be mapped to our familiar V-model or validation lifecycle – there are just a few extra steps and iterations mainly around data and model training. In this part, I'll break down the AI lifecycle into a series of steps and relate each to validation activities. This will give you a blueprint of *what needs to happen* from the idea stage all the way to routine use and retirement of an AI system.

For clarity, I'll structure the lifecycle in **5 major stages** (with some sub-steps), aligning roughly with GAMP's concept→project→operation phases:

Step 1: Concept – Identify Need and Feasibility

Every project starts with an idea or a problem to solve. In this step, you (and your team) identify where AI could be applied and decide if it makes sense.

- **Define the Problem Statement:** Clearly articulate what problem you want the AI to solve or what task to improve. For example, "We want an AI to predict final tablet potency based on real-time sensor data, to adjust the process proactively," or "We need an image recognition system to automate visual inspection of vials." It's important to frame it in business/user terms – that becomes your high-level user requirement.

- **Assess Feasibility:** Not every problem is a good fit for AI. Is the outcome something that can be learned from data? Do you have (or can you get) enough data of good quality to train a model? At this stage, a data scientist might do a quick exploration of available datasets. From a validation perspective, I'm also considering feasibility questions: Is this going to be GMP critical? If yes, what are the regulatory requirements? (Here's where I recall, "if it's critical, no online-learning black boxes!" per Annex 22).

- **Form the Team:** As recommended by regulators, involve a cross-functional team from the start. In concept phase, that might mean QA/validation, process SMEs, IT, and a data science expert discussing the idea. This ensures everyone understands the scope and the risks from their perspective.

- **Initial Risk Brainstorm:** Even at concept stage, it's good to brainstorm high-level risks. For instance, if considering an AI for a critical quality decision, a known risk is "What if the model is wrong and we release bad product?" List these concerns early – you'll formally assess them later, but early awareness will shape the project plan (maybe you

realize you must keep a human decision step in the process – making it a decision support tool rather than a fully autonomous decision maker).

Validation deliverables from Step 1:

- *User Requirements (URS):* Draft the URS to include specific performance expectations for the AI (e.g., "The model shall predict potency within ±5% error for 95% of batches").
- *Feasibility Report (optional):* If an experiment was done on sample data, document those results.
- *Validation Plan draft:* At concept stage, I often start a rough validation plan outlining what stages and special tests might be needed if we proceed. This isn't required by regulations per se, but it helps to foresee effort and is useful when presenting the project to QA for buy-in.

Step 2: Data Preparation – Gather and Curate Data

Data is the fuel of AI. In this step, the focus is on obtaining and preparing the data that will train and test the model.

- **Data Collection:** Identify sources of data. In GMP, data could come from historical batch records, laboratory information management systems (LIMS), sensors, images, etc. For example, if building an AI for visual inspection, you might gather thousands of past product images labeled as "defective" or "okay" by QC. Make sure to consider data representing all relevant scenarios (various products, shifts, machines, etc.). EMA's guidance suggests a *"comprehensive characterization of the data"* needed for the model (health.ec.europa.eu) – meaning think of all input variants the model should handle.
- **Data Cleaning:** Real-world data is messy. You'll likely need to clean it – fix errors, handle missing values, normalize formats. For GMP, ensure you don't inadvertently tamper with raw data without traceability. It's good practice to archive raw data and document any transformations applied (for audit trail purposes and in case you need to defend data integrity).
- **Data Labeling:** If your AI is doing a prediction or classification, you need *labeled* examples (for supervised learning). For instance, to train an AI to detect contamination in cultures, you need plates labeled "contaminated" or "clean" by experts. If labels come from humans, define a procedure for how labeling was done and verified (Annex 22 expects high confidence in reference labels of test data. Sometimes a second independent reviewer verifies a sample of labels to ensure accuracy.
- **Split Data into Sets:** This is crucial. Typically, you split into:
    - **Training Set:** for model training.

  - **Validation Set (optional):** sometimes used for tuning model hyper parameters during development.
  - **Test Set:** held-out data for final evaluation of the model. Annex 22 calls this out as "Test Data Independency" – the test set should be separate and *not* used in training (health.ec.europa.eu). Think of the test set as analogous to the PQ protocol: it should challenge the model with new data to see if it meets requirements. Ensure the split is done in a way that prevents leakage (no overlapping data points). Document how you split it – e.g., "Out of 10,000 images, 8,000 were randomly chosen for training, 2,000 reserved for testing, ensuring that images from the same batch are not split across train/test to avoid bias."

- **Data Augmentation (if needed):** Sometimes, especially in image or signal processing, you might augment the training data (e.g., rotate images, add noise) to simulate more examples. If you do this, document it. Augmentation can improve model robustness, but be cautious: augmentation should represent realistic variations.

Validation deliverables from Step 2:

- *Data Management Plan / Data Specification:* I create a document describing all data sources, how data was verified or cleaned, how it's stored, and the rationale for data sufficiency. Also note if data is considered *GMP raw data* or not (often training data might not be traditional GMP data, but test data might come from GMP records – so consider integrity).
- *Traceability of Data:* Sometimes we keep a list of input data files or a log of data extraction queries to show traceability from source systems.
- No formal testing yet, but this stage sets you up for the model building. Review this plan with QA if possible – it's new for many QA folks, so walking them through data prep steps helps them get comfortable that you're treating data carefully.

Step 3: Model Development – Build and Train the Model

Now the data scientists (or developers) get to work creating the AI model. This stage is typically iterative – try a model, see results, adjust – but let's break it into sub-steps:

- **Select Algorithm/Model Type:** Based on the problem and data, choose an approach. It could be a classical ML algorithm (like a decision tree, regression, etc.) or a complex one (like a deep neural network). From a validation perspective, complexity matters: simpler models are easier to validate and often more interpretable. If a simpler model meets the needs, that might be preferable in GMP. If a complex model is chosen, be ready to justify why (perhaps the problem is too complex for simpler methods).

- **Develop Model Code:** Write or configure the model. In many cases, you'll use existing libraries (TensorFlow, scikit-learn, etc.). This is like the configuration/customization in a regular system project. Ensure the development environment is controlled – e.g., you know which version of libraries were used to train the model (because if you retrain with a different version, results could differ). GAMP recommends leveraging supplier documentation if using off-the-shelf algorithms, but in all cases, **maintain version control** on your model code and training scripts – they are part of the software configuration.
- **Train the Model:** Run the algorithm on the training data to produce a trained model. This can be computationally heavy and might be done by data scientists outside the validated production environment (often it's done in a research or development environment). That's okay – model training is a development activity. You'll later deploy the final model into a validated environment. Just make sure the final model's parameters (weights, etc.) are saved and under version control as a configuration item. In GMP terms, think of the trained model file like an installation package – you will install that into production.
- **Evaluate on Validation Set (if using one):** During tuning, the data scientist might check how the model does on a validation set, tweak hyper parameters (like learning rate, number of layers, etc.) and retrain to improve. This is akin to unit testing or iterative development in traditional terms.
- **Ensure No Peek at Test Set:** It's tempting to see how you're doing on the test set, but resist using it until you're ready for final evaluation. This is like not using your PQ test cases during development – you want an unbiased final test.
- **Documentation During Development:** Encourage the team to document as they go. For instance, keep notes on different model versions tried, what their performance was, and why the final model was selected. This can go into a Model Development Report for transparency. Also, log random seeds or any factors to ensure reproducibility (if someone retrains the model with the same data, do they get a similar result?).

At the end of this step, you should have a *candidate model* that you believe meets the requirements. For example, you might have a neural network model file that you think can identify defects with 98% accuracy based on internal testing.

Validation deliverables from Step 3:

- *Model Design Specification:* Sometimes I write a short spec describing the model architecture (e.g., "a 3-layer neural network with X inputs…") and how it works. This is helpful for future maintainers and for QA to understand what we have built.
- *Model Training Report:* Detailing training runs, parameters, and internal evaluation results. This is evidence that due diligence was done and can be appended to the validation package. It might include plots of training vs. validation accuracy, etc.
- *Config Management:* Check that the final model (and training code) are checked into a configuration management system (with an identifier or version number). Also note the environment (software libraries, versions) used.
- We haven't done the formal "validation testing" yet, but internal results are documented here.

Step 4: Verification & Validation – Test the Model and System

Now comes the moment of truth: formal testing of the model and the overall system in which it operates. This maps to our traditional IQ/OQ/PQ phases, adapted for AI:

- **Installation Qualification (IQ):** If the AI solution involves installing software (e.g., deploying the model into a production application or device), do an IQ to verify installation parameters. For example, if you have an AI-enabled instrument, IQ would ensure the hardware/software is properly installed, the correct model file (the one we validated) is deployed, and all dependencies are satisfied. IQ might also involve verifying that the computing environment (server, cloud, etc.) meets security and infrastructure requirements (especially if the AI is deployed on a new platform).
- **Operational Qualification (OQ) – Model Performance Testing:** This is where we use the reserved **test dataset** to challenge the model and see if it meets the acceptance criteria defined earlier. It's essentially a scientific experiment, executed under controlled conditions:
    - You feed the model the test data inputs and record the outputs (predictions).
    - Compare the outputs to the known expected results (the "ground truth" labels).
    - Calculate the performance metrics (accuracy, sensitivity, etc.) and see if they meet the pre-set acceptance criteria. For example, "Out of 500 defect images, the model detected 490 (98%), which meets the ≥95% criterion. False positive rate was 0.5%, meeting the ≤1% criterion."
    - Don't forget edge cases: ensure the test set included some worst-case scenarios. If any specific failures were anticipated (like certain rare defect types), check how the model did on those.

  - Document any deviations: if the model fails a criterion, that's a deviation. At this point, you have options: if it's minor, you might justify an acceptance (risk assessment, maybe the criterion was too strict). Often though, a failure means loop back to development – adjust model or get more data, then re-test. Regulators allow this iterative approach, but you must document it (just like any validation deviation and re-test).
  - **Explainability & Robustness Tests:** Beyond basic metrics, sometimes you perform additional tests. For example, slightly perturb inputs to see if the model output changes dramatically (to test robustness), or use explainability tools to confirm the model is looking at sensible features (this is not always mandatory, but a good practice especially if trying to convince an inspector that the model's behavior is reasonable).

- **OQ – Functional Testing of Surrounding System:** If the AI is part of a larger system (almost always the case), test the integration. For instance, if the model is deployed in a software that laboratory analysts use, test that workflow: Does the system properly load the model, feed it data, and display the AI's output? If the AI flags a result, does it generate the correct alert? Essentially, test all functions including those with the AI in the loop and those without. Treat the AI like any other component: challenge any error-handling, ensure it fails safely if, say, input is out of the expected range (Annex 22 requires monitoring if input goes out of the trained scope (health.ec.europa.euhealth.ec.europa.eu).
- **Performance Qualification (PQ):** In many AI validations, the distinction between OQ and PQ can blur, but here's one way to see it: OQ might be done in a controlled test environment with historical data, while PQ could be a *prospective* trial of the system in actual operation. For example, you might run the AI system in parallel with the existing process for a period of time (without it actually controlling anything) to confirm it performs well in the real production environment with live data and users. During PQ, you gather user feedback too: are the lab analysts comfortable with the AI tool, do they understand its output, is the response time acceptable, etc.? This ties into the *human factors* – an often overlooked aspect. A tool is only effective if people use it correctly.
- **Final Risk Assessment:** After testing, update the risk assessment. Have all identified risks been mitigated or accepted? For instance, if a risk was "model might not generalize to new product variant," and during PQ you included some new variant data and it worked, great – risk mitigated. If not, you might keep that risk open and plan how to address it (maybe restrict use of AI to certain product types until retrained).
- **Qualification of Users: I**f the AI tool changes the way people work, ensure training is done. For example, if it's a decision support system, train the operators on how to

interpret the AI output and what to do if they suspect the AI is wrong. (This is part of operational readiness and also a control for risk of over-reliance.)

Validation deliverables from Step 4:

- *IQ Protocol/Report:* Verify installation, environment, configuration (including verifying you're using the correct final model file – I often do an MD5 checksum or similar to ensure the model in production is exactly the one we tested).
- *OQ/PQ Protocols:* Detailing test cases, especially the model performance test. Each test case might correspond to a metric or scenario (e.g., "Test Case 1: Model accuracy on known good vs bad samples – expected ≥95% detection of bad samples").
- *Test Results and Report:* Document actual results and whether acceptance criteria met. Include summary tables of performance.
- *Deviation Reports:* If any deviations occurred and how resolved.
- *Final Validation Report:* Sum up that the system (including the AI) is validated for the intended use, referencing all evidence. It should explicitly state any limitations (for instance, "Model validated for products A, B, C. Use with any other product will require further validation.").
- *SOPs and Training Records:* Ensure SOPs are updated for operating the system, including what operators or QA should do with the AI outputs. For example, an SOP might say "If AI predicts a batch quality issue, QA must review the evidence and decide on batch disposition. AI prediction alone is not a batch release criteria." This kind of instruction ensures compliance and aligns with the concept of keeping a human in the loop for critical decisions.

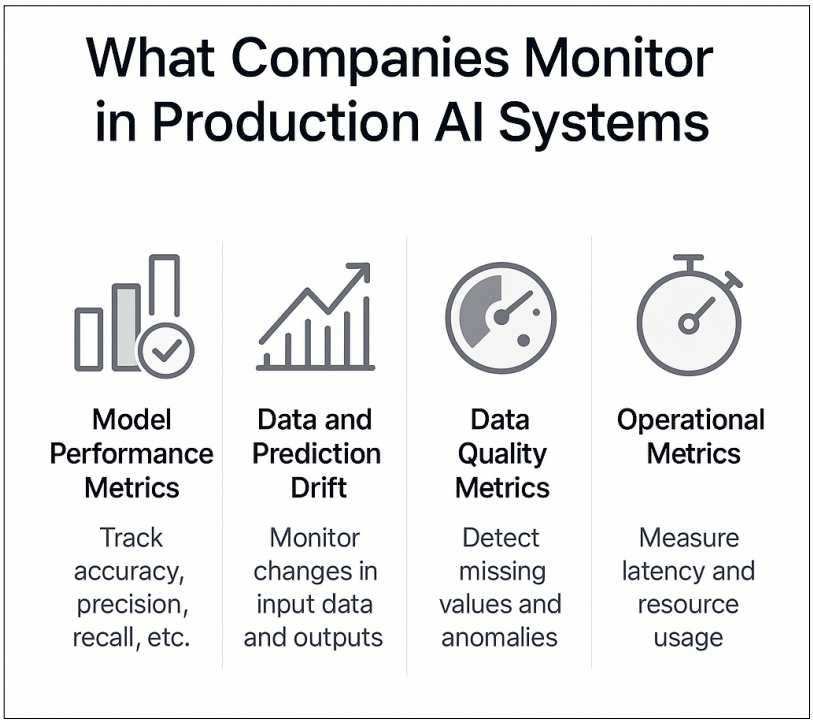
Step 5: Deployment and Ongoing Monitoring – Maintain the System

Validation isn't a one-and-done deal, especially for AI. Once the system is live, GMP expectations (and good sense) require that we **monitor and maintain** the validated state. This final stage is about running the AI system in production and ensuring it continues to perform over time.

- **Go-Live (Release for Use):** After successful testing and QA approval, formally release the AI system for operational use. This might involve moving the validated model and software into the production environment if testing was in a staging area.
- **Monitor Performance Metrics in Production:** Set up a mechanism to routinely check the AI's outputs. For instance:
    - If it's making predictions continuously (like a process control prediction each batch), track how often those predictions are correct by later comparing to actual outcomes. This is like continued process verification for the AI.

- If it's a lab tool, maybe have periodic quality control samples with known outcomes run through it to ensure it still catches what it should.
- Many AI systems can log confidence scores or internal statistics – keep an eye on those. Annex 22 suggests regular monitoring for any *drift* in input data or model performance (health.ec.europa.euhealth.ec.europa.eu). So if you start seeing inputs that are outside the ranges seen in training (data drift) or if outcomes start deviating, that's a trigger to investigate. Refer to Appendix B for more monitoring details.

- **Retesting or Recalibration Schedule:** Consider if the model needs periodic re-validation. For example, "We will re-evaluate the model annually using a fresh set of data from the last year to ensure performance hasn't degraded." If performance is dropping, you may need to retrain the model with new data (which would be a controlled change).
- **Change Control for Model Updates:** Any planned change, like updating the model (retraining it on more recent data to improve it) or upgrading the algorithm, must go through change control. This includes an impact assessment (e.g., "retraining with 20% more data including new product – will extend model to cover product D") and requires re-validation testing. FDA's concept of a "Predetermined Change Control Plan" for AI in devices (although different context) is relevant: plan how you'll update models and what tests will be needed (FDA has guidance on this for adaptive AI algorithms). In GMP, we likely won't have autonomous updates – you'll always treat a new model as a change request that QA must approve after evidence.
- **Incident Management:** If the AI system ever has an unexpected behavior (e.g., it gave a wildly incorrect recommendation that a user caught, or it wasn't available when needed), log it as you would any deviation or incident. Investigate root cause (was there an unanticipated scenario? a data pipeline error? etc.), and take corrective action (maybe add that scenario to the training set next version, or adjust thresholds, or improve user training).



- **Periodic Review:** It's good practice to periodically review the whole system (perhaps as part of the periodic IT system review). Check if the regulatory landscape changed (new guidelines?), if the process context changed, if any new risks emerged. Also, ensure documentation (SOPs, etc.) is current with any minor updates.
- **Retirement Planning:** Eventually, if the AI system is replaced or retired, ensure proper retirement procedures – archive the model, data, and code as needed (so you have traceability in the future), and verify it's removed from active use.

By maintaining vigilance in the operation phase, you keep the system in a validated state and avoid nasty surprises. Regulators will expect to see evidence of this monitoring. For example, during an inspection, an inspector might ask, "Show me how you know the AI is still performing as intended since you validated it a year ago." You could show them a trend chart of model accuracy over each batch for the past year, or summary reports of periodic re-validation tests. This demonstrates *continuous state of control*, which is the essence of process validation and applies equally to AI models in GMP (appendix B).

In summary, the AI lifecycle has many familiar elements (requirements, testing, change control) with a few added twists (data prep, model training). Understanding these steps helps structure our validation project plan. Now that we've laid out the lifecycle, the next section will dive deeper into **risk management** – which underpins decisions at every stage (from choosing where to use AI, to setting acceptance criteria, to deciding on controls).

Let's move on to how to think about and manage risks when dealing with AI in a GMP context.

5. Risk Management for AI in GMP

If there's one theme that comes up repeatedly in all guidances for AI in regulated settings, it's **risk management.** As a CSV professional, you're already familiar with the concept: identify what could go wrong, how likely and severe that is, and put controls to prevent or mitigate it. For AI systems, we apply the same *Quality Risk Management (QRM)*principles (like ICH Q9) – but we need to account for some new risk factors introduced by AI's nature. In this section, I'll outline a risk management approach tailored to AI in GMP, including typical risks to consider and how to address them.

Unique Risk Factors of AI Systems

First, let's enumerate what new or heightened risks AI brings compared to traditional systems:

- **Model Inaccuracy or Failure:** The model could make wrong predictions or decisions. This is obvious, but in GMP the key is the *impact* of a wrong prediction. Does a false prediction lead to a minor inconvenience or a major quality risk? For example, an AI that misflags a good batch as bad (false positive) might cause a delay or investigation (costly, but no patient harm), whereas missing a bad batch (false negative) could release a substandard product (patient risk).

- **Lack of Determinism:** Traditional software gives the same output for the same input, and if it doesn't, that's a bug. Some AI models (especially if not constrained as per Annex 22) could behave non-deterministically or evolve. Even deterministic models might behave unpredictably if fed data outside what they were trained on. This unpredictability is a risk – we mitigate it by restricting AI use to defined domains and including human review for critical decisions.
- **Data Quality and Bias:** If the training data was incomplete or biased, the model will carry that bias. For instance, if all training data for a visual inspection AI came from one production line, the AI might not perform well on another line with slightly different lighting or camera calibration. Or more subtly, if the model is used in a lab to identify atypical results, but all training data came from experienced operators, the model might not catch issues that a newbie operator's technique might introduce (because it never saw those in training). Data bias can lead to systematically skewed decisions.
- **Cybersecurity and Data Integrity:** AI models could be targets for tampering. It's a newer area, but consider "adversarial examples" (someone intentionally feeding inputs that fool the model) or hacking the model to alter its weights. In GMP, this might not be a common scenario, but we should still ensure the model file and inputs are secured so that no one can maliciously or accidentally alter how the AI behaves. Treat the model as a piece of software that needs access control and checksums (like any executable).
- **Lack of Explainability (Black Box):** If we can't easily explain why the AI made a certain decision, it's harder to trust and harder to debug when something goes wrong. This is a regulatory concern as well as a practical one. It's a risk if users are asked to rely on an AI output without understanding it – they might over-rely (thinking it must be right) or under-rely (ignoring it completely). Both situations reduce the effectiveness of the system. We mitigate this by providing context or rationale for AI outputs when possible, or at least by educating users about the model's limitations.
- **Model Drift Over Time:** The model's performance could degrade as real-world conditions change. For example, an AI model for environmental monitoring anomaly detection might gradually become less accurate if the typical environmental data changes (maybe due to new equipment or seasonal variation) – this is *data drift*. Or if it's a predictive maintenance AI on equipment, and the equipment ages or is modified, the old model might not be as predictive – *concept drift*. This risk is mitigated by continuous monitoring and periodic retraining (refer to Appendix F).
- **Human Factors:** Ironically, one of the biggest risks can be *us*, the humans. How operators and decision-makers interact with AI matters:
    - **Over-reliance risk:** People might trust the AI too much and not catch its mistakes (especially if it has performed well in the past). For example, if an AI decision support has never been wrong in a year, operators might start rubber-stamping its recommendations without due diligence.

- **Disuse or Workaround risk:** If the AI system is cumbersome or occasionally wrong, people might ignore it or find ways to work around it, defeating the purpose (like always overriding the AI, or feeding it data that biases it to be conservative so they feel safer).
- Proper training and clear SOPs can mitigate these: make sure users know the AI is a tool, not an oracle, and that they have defined responsibilities to review or verify certain things.

Applying QRM (ICH Q9) to AI

To manage these risks, I use a standard QRM process: Risk Assessment (identify, analyze, evaluate), Risk Control (reduce or accept), and Risk Review (monitor over time). Let's go through an example of how this might look for an AI system:

Risk Assessment:

1. **Identify Hazards:** Gather your cross-functional team and brainstorm what could go wrong. Use prompts like "What if the model is wrong? What if we feed it bad data? What if the system is unavailable? What if a user misinterprets an output?" Be specific: e.g., "Model falsely predicts a good batch is bad and batch is needlessly scrapped" or "Model misses a contaminant in a vial and batch released with defect." Also consider regulatory compliance risks: "What if we cannot explain a decision to an auditor?" or "What if the model update process fails and an unvalidated model is deployed inadvertently?"
2. **Analyze Risk:** For each identified scenario, estimate severity, probability, and detectability (typical FMEA-style approach). For instance, missing a contaminated batch (severity high – patient safety issue; probability maybe low if model is generally good; detectability maybe low if no human double-check – hence risk is high). By contrast, a false alarm causing a batch hold (severity medium or low; probability maybe moderate; detectability high because investigation would find it was false – hence risk is moderate or low).
    - Pay attention to how existing controls factor in. If there is a human verification step, that improves detectability of model errors. If data goes through verification (like an analyst reviews images the AI flagged *and* those it didn't flag maybe in spot-checks), that's a control lowering risk.
    - We might quantitatively use risk scoring, but qualitative is fine. The outcome should be a prioritized list of what we must absolutely manage versus what is minor.
3. **Evaluate Risk Acceptability:** Decide which risks are unacceptable and must be mitigated. Generally, any risk impacting patient safety, product quality, or data integrity in a significant way needs mitigation. Lower ones might be accepted with rationale. For AI, regulators expect a conservative stance: e.g., the reflection paper suggests if an AI tool impacts the benefit-risk of a medicine, involve regulators earlyema.europa.eu – implying such a tool's risks better be well controlled.

Risk Control:

Now for each major risk, plan controls. Controls can be **preventive** (stop the error from happening) or **mitigative** (catch it before harm).

Let's go through common risks and typical controls for an AI system:

- *Risk:* **Model makes an incorrect critical prediction (false negative or false positive).**
  Controls:
  - Preventive: Use a high-quality model (by thorough validation and choosing a robust algorithm). Set conservative decision thresholds (e.g., if model's confidence is low or data is out-of-scope, have it default to "unable to decide" and escalate to human). Also, you might design the process such that the AI assists rather than finalizes critical decisions (so a human is always there as a safety net).
  - Mitigative: Implement human-in-the-loop review. For example, require QA to review all batch release recommendations from the AI, or have an operator confirm any vial rejects flagged by AI. Essentially, someone can catch the AI's mistake. Also, monitor model performance indicators for any sign of degradation (so you catch if it starts being wrong more often).
- *Risk:* **Model not performing well on a new scenario (data drift or new product).**
  Controls:
  - Preventive: As part of change control for introducing a new product or significant process change, require an assessment of the AI model. If new inputs are outside its original training scope, retrain or at least test the model on some data from the new scenario before fully trusting it.
  - Mitigative: Continuously monitor input data distribution. Some teams set up automated checks – e.g., using statistical tests to see if current process data significantly differs from training data. If yes, trigger an alert that model may need retraining. Annex 22 explicitly expects monitoring whether input data is still within the model's "sample space"health.ec.europa.eu.
- *Risk:* **Data integrity issues (feeding AI bad data unknowingly).**
  Controls:
  - Preventive: Put in data validation checks in the data pipeline. For example, if an instrument feeds data to the AI, ensure there are sanity checks (no negative values where impossible, flags if data is incomplete). Use only validated sources of data. If data comes from a manual entry, apply usual data integrity controls (double data entry or review).
  - Mitigative: If AI output seems outlandish, have procedures that prompt a manual data quality check. Often, an extreme AI result can indicate a data problem (like a sensor calibration issue).
- *Risk:* **Model bias leads to systematic error (e.g., always under-predicts for a certain product).**
  Controls:

  - Preventive: During development, include diverse training data. Also, test the model separately on key subgroups of data to see if there's bias (Annex 22 suggests performance may be evaluated per data subgroup health.ec.europa.eu). If bias is found, retrain with more balanced data or adjust the model.
  - Mitigative: Operationally, track metrics per subgroup or context. For instance, track prediction error per product type; if one product consistently has worse predictions, that could indicate bias that needs correction.

- *Risk:* **User misinterprets or misuses AI output.**
  Controls:
  - Preventive: Provide clear user interface and guidance. For example, if an AI gives a risk score 0–1, label it with categories ("Low/Med/High") to reduce confusion. Also, train users on examples of how the AI works, including its known limitations. People need to know, *for example, "This AI is not 100% accurate, if something seems odd, investigate further."* Include that in SOPs: maybe require a routine audit of some percentage of AI decisions by a second person.
  - Mitigative: If feasible, have periodic review of decisions made with AI assistance. For instance, QA might do a monthly audit of a few decisions to ensure the operator correctly followed up on AI recommendations and didn't ignore critical ones.

- *Risk:* **Technical failure (system goes down or model doesn't run).**
  Controls:
  - Preventive: Standard IT controls – robust infrastructure, backups, fail-safes. If the AI is critical, have a fallback process (e.g., if AI is unavailable, revert to manual process). Validate that fallback process too or at least define it.
  - Mitigative: Alarms for system failure, so that if the AI system fails, it's noticed immediately and addressed, and the manufacturing doesn't blindly proceed without its checks (or conversely, doesn't grind to a halt unnecessarily).

- *Risk:* **Regulatory non-compliance (unable to explain or justify system during audit).**
  Controls:
  - Preventive: Maintain thorough documentation as described throughout this guide. Have clear rationales for why the model is acceptable, and reference guidances (e.g., document how your approach aligns with FDA/EMA guidance – which makes an auditor's job easier to see you did your homework).
  - Mitigative: Ensure someone in the company (maybe you, maybe a data scientist) can speak to the details if needed. Conduct internal audits of the AI system and its validation to prepare for questions.

We should also assign **risk owners** for ongoing risks. For example, after deployment, who is responsible to monitor data drift? It could be the process owner or a data scientist in QA. Assign it so it doesn't fall through the cracks.

Risk Review:

Risk management isn't one-time. Periodically (say annually, or whenever a major change occurs), review the risk log:

- Check if controls are effective (has any risk manifested despite controls? E.g., any incident of AI error? If yes, update the risk assessment with that occurrence and what was done).
- Update risk analysis if new risks appear (maybe after some experience with the system you discover a new failure mode).
- Feed this back into any future improvements or retraining of the model.

Regulators appreciate seeing that you treat the AI system with the same vigilance as any critical process – that means doing formal risk assessments and keeping them live. For instance, the EMA reflection paper notes that *applying ICH Q9 QRM principles from early development through deployment* is key for AI. Also, PIC/S/EMA explicitly integrated risk management into Annex 22 (section 2.3 of Annex 22 says activities should be based on risk to patient safety, product quality, data integrity (health.ec.europa.eu). So by documenting your QRM steps, you're directly addressing that expectation.

One tool I've used is a specialized **FMEA for AI.** It's like a spreadsheet where each potential failure of the AI (failure mode) is listed, cause analyzed (like cause could be "inadequate training data for scenario X"), current controls listed, and risk scored. It helps structure what I described above. If the risk scoring is high, you add additional controls or decide not to use AI for that scenario at all.

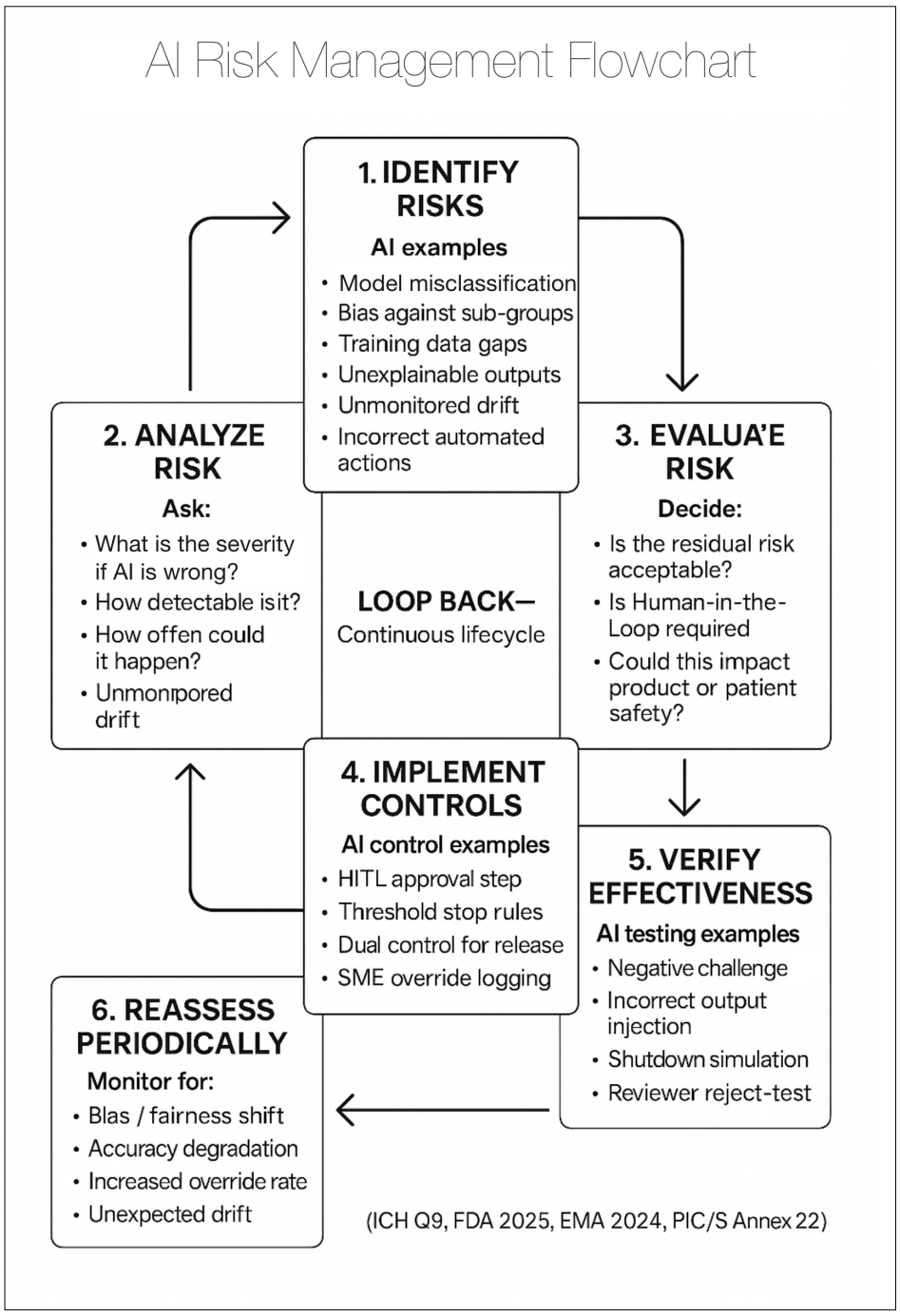
For example, if risk of "model missing a critical contaminant" remains unacceptably high even after controls, you might decide: this AI will only be used as a secondary check, not the primary method, until it's improved. That is a valid outcome – risk management may lead you to scope down the AI's role.

Documenting Risk Management: Usually, I produce a *Risk Management Plan* early (stating we will do QRM and what methods) and a *Risk Assessment Report* that records the details (or include it within the validation plan/report). This should be referenced in validation docs and also in SOPs if needed (like "based on risk assessment, this system is considered high risk and thus will be reviewed quarterly by QA" – making that a formal practice).

To sum up, **risk management for AI is about being proactive and vigilant.** We accept that AI isn't perfect, but we make sure that when it's not perfect, it doesn't harm patients or product quality. By identifying failure modes and building in safeguards (like human oversight, alerts, periodic checks), we can reap the efficiency or insight benefits of AI while still protecting what

matters in GMP. And beyond compliance, this gives everyone (from management to shop floor operators to auditors) confidence that the AI is a help, not a potential lurking danger.

With risk management strategies in mind, we're well-equipped to design a robust validation approach. In the next section, let's get very practical: how do we apply all this to actual *use cases* of AI in GMP, like decision support tools and lab automation systems? We'll walk through examples and validation tips for those scenarios.



6. Validation Framework for Decision Support Tools and Lab Automation

Now for the fun part: putting it all together for real-world examples. Many of us will first encounter AI in the form of **decision support tools or lab automation systems.** These are areas where AI can add a lot of value – by analyzing complex data to support human decisions, or by automating tedious lab tasks with smart algorithms – but they also need to be validated and controlled like any GMP system.

In this section, I'll share a framework (and tips) for validating these two common categories of AI applications. We'll go through an example use case for each, highlighting what to include in validation deliverables, test plans, and procedural controls. I'll continue in a first-person narrative as if I'm walking you through a project I handled, as I find that more engaging and practical.

Validating an AI Decision Support Tool (Use Case Example)

Use Case: Let's say we implement an AI-driven **Deviation Management Assistant** in our quality unit. The idea is the AI will review deviation reports (manufacturing deviations, lab OOS reports, etc.) and predict which ones are likely critical, which are likely minor, and even suggest potential root causes or similar past cases. The goal is to help QA triage and investigate deviations faster by focusing on the most risky ones first and learning from history.

Crucially, this is a **decision support** tool – it does not make the final decision on classification or CAPA; it provides a recommendation to the QA specialist, who then uses that input plus their judgment to proceed.

Planning Validation:

1. **Requirements:** We gather requirements like:
    - It should accurately classify historical deviations: e.g., at least 90% of deviations that were ultimately classified as critical should be flagged as high-risk by the AI.
    - The AI should provide an explanation or reference (like "flagged because similar deviation X last year was critical due to root cause Y").
    - Response time, integration requirements (it should integrate with our electronic QMS software).
    - Importantly, define what happens if it's wrong. Since this is advisory, a wrong advice should not automatically harm – but we require that it never auto-closes a case or anything. That's a requirement: system shall not automatically disposition a record without human confirmation.
    - Also, maybe a requirement: if the AI is not confident (say <50%), it should indicate uncertainty rather than give a potentially misleading guess.

- We also include usability requirements, like it should rank deviations by risk, have a user-friendly interface, and log all recommendations given.

2. **Risk Assessment Highlights:**
   - We identify that a major risk is *over-reliance*: QA might blindly trust the AI and potentially miss a critical deviation if the AI failed to flag it. So we plan a control: no matter what the AI says, QA's SOP says "you must still review all deviations, AI is just a tool." That way, a critical deviation should still be caught by human review even if AI misses it. AI is just prioritizing, not eliminating work.
   - Conversely, false positives (flagging many things as critical) might cause alarm fatigue. So we set a performance requirement to minimize false alarms to a reasonable level and plan to monitor that.
   - Data for this AI: it will be trained on our deviation database from past 5 years. Risk: if our historical data labeling was inconsistent, the AI might learn a bias. We mitigate by cleaning that data as well as possible (maybe only using deviations that had final QA approval on classification).
   - Regulatory risk: We must be careful that this tool doesn't inadvertently bypass any GMP decision-making requirements. We involve QA management in reviewing its role, and likely update procedures to formalize how it's used (ensuring it's an aid, not a decision-maker).
3. **Development and Testing:**
   - We train the model (perhaps a natural language processing model combined with some rules) on historical deviations.
   - We hold out last year's deviations as a test set to simulate how it works on "new" data.
   - The team tunes it until it can predict critical vs minor with, say, 95% accuracy on the test data. It also suggests a likely cause category correctly say 85% of time.
   - We implement it into the QMS as a separate module that reads deviation text and outputs a risk score and suggestions.
4. **Validation Execution:**
   - **IQ:** Verify the module is installed in QMS test environment, correct model version loaded.
   - **OQ:** Use a test data set (could be the hold-out or a curated set of example deviations with known outcomes) to formally test:
     - Does it flag the right ones as high risk? (We compare to known classifications.)

- Does it correctly pull up similar past cases? (We have expected examples – e.g., we know deviation 123 should bring up deviation 45 from last year as similar – test that).
- Does it respond with uncertainty when appropriate? (Maybe craft a deviation that is completely new type – expect AI says "unsure").
- Also test integration: e.g., when a new deviation is entered, does the system automatically produce a recommendation? Test the timing and logging.
- Test user interface: ensure the recommendation is clearly displayed and labeled as "AI-generated suggestion" to avoid confusion.
- Test failure modes: e.g., what if the AI service is down – does the QMS just show "no suggestion available" and allow QA to proceed manually? (This needs to be acceptable).

- **PQ:** Possibly run a pilot – let's say for one month, QA uses the tool but double-checks everything. We gather feedback:
  - Did it actually help? Did QA find it accurate? This is subjective but important – if users find it consistently unhelpful or confusing, that's a process risk.
  - We also track if any critical deviation was mis-prioritized during the pilot (none should escape QA review because QA still looks at all, but was any critical one ranked low by AI? If yes, we analyze why and possibly improve model).
  - If pilot is good, we go live fully.
- Throughout OQ/PQ, we keep evidence. Let's say out of 50 test deviations, AI correctly classified 47 – that's 94%, meeting our 90% criterion (pass). It suggested a correct root cause category in 40 of 50 (80%) – slightly below our 85% target, maybe we discuss whether to accept or retrain. This is a judgment call; maybe 80% is still a huge improvement over nothing, and the suggestions are only aids. We might accept 80% with a note that it will improve as more data is added and we'll retrain next year.
- We also do a user acceptance test: a couple of QA folks use a test instance to make sure they understand the outputs and the interface is fine.

5. **SOPs and Training:**
   - We update the Deviation handling SOP: "QA Specialist may utilize the AI Deviation Management Assistant as follows: ... However, the final classification

and investigation decisions must be made per QA's judgment. All AI suggestions are to be reviewed and, if overridden, document reason..." etc.

- We train QA specialists on the tool – how to interpret the risk score and what to do if they disagree with it.
- We also set an SOP for model maintenance: e.g., "Quality Systems department will retrain the model annually with the latest data and perform a re-validation of performance. Any retraining must be approved via change control."

6. **Monitoring:**
    - We plan to monitor: e.g., every quarter, sample 10 deviations and see if any that ended up critical were initially not flagged by AI – report that as a metric.
    - Also monitor if AI flagged many things that turned out not critical – if it's too many, maybe adjust threshold or retrain.

Documentation: The validation package for this includes URS, Functional Specification (maybe describing algorithm logic), Risk Assessment (with focus on the AI risks and controls as above), IQ/OQ/PQ protocols and reports, and the SOP updates. We include references to EMA reflection paper statements supporting why we keep a human in loop (e.g., human-centric approach) and Annex 22 points (though this is not manufacturing, the principles still apply: intended use clearly defined, metrics etc.). For instance, we can cite that *regulators require human oversight for AI outputs in GMP (*health.ec.europa.euhealth.ec.europa.eu), which backs our SOP control.

This example highlights that for decision support tools:

- **Keep the human in control:** We treat AI as an aid. This generally lowers risk and validation burden because the ultimate decision remains with a qualified person. In terms of validation evidence, we focus on demonstrating the AI's recommendations are reasonably reliable and definitely beneficial.
- **Focus on integration and user interaction:** The validation must ensure the recommendation gets to the user clearly and is used appropriately. This is less of a concern in pure automated systems but huge in decision support. So user testing and SOP alignment are part of validation.
- **Allow for AI fallibility:** We explicitly plan for what happens if AI is wrong or uncertain (via process design, which is a form of risk control).

Validating an AI-Powered Lab Automation System (Use Case Example)

Use Case: Suppose we introduce an AI-based **Automated Colony Counter** in our microbiology lab. Traditionally, lab analysts count bacterial colonies on plates manually or with simple image software. Our new system uses a machine learning vision algorithm to count colonies and even distinguish different morphology (say it can identify two different species by color/shape if plates have mixed cultures). This system will automatically generate

the colony count results that feed into the test record. Analysts will review the results, but they won't have to count by hand unless something looks off.

This crosses into lab automation: it's actually making an autonomous measurement (colony count) that goes into a quality record, albeit with review. It's also decision support in a sense (identifying colony types), but let's treat it as automation of a lab test process.

Validation Approach:

1. **System Description and Requirements:**
    - This likely involves a specific instrument (a plate scanner) and a software with AI. We'll gather URS:
        - Accuracy requirements: e.g., "The system shall count colonies with an accuracy of ±1 colony or 5% (whichever larger) compared to manual expert counts, across a range of 0 to 300 colonies per plate." This is based on compendial expectations (if any) or just acceptable error that won't impact product decisions.
        - It shall correctly differentiate red vs white colonies (for example) 95% of the time (if we need species identification).
        - Throughput: can process X plates per hour, etc.
        - Data integrity: It must record images and results in a way that analysts can verify (like storing plate images with the counted spots marked, so a human can audit if needed).
        - Integration: results must feed to LIMS or be stored with sample ID etc.
        - We also include requirement that an analyst can override or correct the count if they see an error (so manual correction feature is needed and that must be audit trailed).
        - Regulatory: since this is effectively performing a test, ensure it complies with data integrity (audit trails of any changes, etc.), 21 CFR Part 11 for electronic records if applicable.
2. **Risk Assessment:**
    - *Risk:* AI miscounts colonies (e.g., conflates two nearby colonies as one, or noise as a colony). Severity depends: a miscount might cause an out-of-spec if it undercounts or overcounts around a spec limit. If the spec is say <100 colonies, and actual is 105 but AI counts 95, that's a potential false pass – high risk. On the other hand, a slight miscount well within spec might have no impact.

- We'll ensure high severity scenarios have mitigations. For example, require any plate that AI counts near the spec limit (within, say, 10% of limit) be flagged for manual review.
- *Risk:* misidentification of colony type – could lead to missing a contaminant species. Mitigation: possibly double-check with confirmatory tests for critical ones, or at least flag any uncertain classification for human ID.
- *Risk:* plate image quality issues (condensation, etc.) – AI might fail or guess. Control: have the system detect if an image is too low quality and alert the analyst to manually inspect/count.
- *Risk:* user blindly accepts count. Control: require periodic verification by analysts. Maybe SOP says each analyst will manually verify a random 5% of plates counted by AI to ensure it's performing.
- *Risk:* data not saved properly – ensure backup of images, etc.
- If the AI is considered part of a standard test method, also consider if we need regulatory method validation (if it's replacing an official compendial method). Possibly not if it's just a different way to count, but one should check guidelines (like USP chapters for automated colony counters).
- Also, dynamic vs static: this model might be static (we likely freeze it). If it's vendor-supplied, maybe it's pre-trained. If we can further train it with our plates, then we should treat any training as a change process.

3. **Vendor Assessment (if applicable):** Many lab AI systems are bought from vendors. We'd audit/document vendor qualifications: Did they develop the model under a quality system? Can they provide validation data? For example, vendor might have a spec "99% accurate on standard colony images". We won't just trust it, but it helps to have their evidence. If vendor is not familiar with GMP validation, we may need to do more ourselves.
    - If we develop in-house (less likely for a lab tool), then all training as per earlier steps.
4. **Installation and Configuration:**
    - Setup the instrument, install software. Here IQ involves checking the correct AI model file version (maybe the vendor provides model v1.2), calibration of the imaging hardware, etc.
    - Also, define user accounts, ensure Part 11 settings (passwords, roles) in software.

5. **Testing (OQ/PQ):**
    - We need a set of **validation plates** to test performance. We might prepare, say, 50 plates with known colony counts (some low counts, some high, some near spec, etc.). How to get "known"? Perhaps have two experienced microbiologists count them manually (the "truth"). Or artificially create spots of known number (like printed reference plates, though real colonies vary).
    - We run these plates through the AI system:
        - Compare AI counts to manual counts (the truth). Calculate accuracy, etc. If any are out of tolerance, investigate plate image – was it something AI missed? This is our formal OQ for counting.
        - Test colony identification function: maybe we have some plates with two species dyed differently. Check confusion matrix (like how many of species A did it label as B, etc.).
        - Test edge cases: a plate with confluent growth (should report "too many to count – TNTC"), a plate with artifacts like a scratch or writing on it – does it ignore those? These challenge tests ensure the system handles or flags anomalies properly.
        - Test the software features: e.g., if an analyst disagrees with a count, can they edit it and does it log the change? Do images get saved and can they be retrieved and re-analyzed if needed?
        - If the system integrates to LIMS, simulate that: ensure results transfer correctly.
    - **Acceptability criteria:** The system might not need 100% match with manual, but we likely set a criteria like "No more than ±5% difference on 90% of plates; no single plate >10% off; and in no case shall a result cross a spec threshold incorrectly." The last part might be tested specifically: if spec is 100, we include some plates around 90-120 colonies and see if any miscounts would have caused a false pass/fail.
    - If any failures: e.g., AI counted 85 vs actual 95 on one plate – that's a 10 difference. If our limit was ±5, it fails. We investigate: maybe that plate had a weird issue. We might adjust the model (if we can retrain or adjust threshold) and re-run, or decide this scenario (maybe colony overlapping) will always be flagged for manual review as a procedural control, then we can accept that and note it.
    - **PQ:** Use the system on real samples in parallel to old method for a while. E.g., count colonies manually and with AI for say 30 routine samples. Check discrepancies. If all good (or any differences well within acceptable range and did not affect decisions), then we have confidence to fully switch.

- During PQ, also gauge user comfort: do analysts trust it? Did it speed them up? Any feedback like the interface annotation could be improved? We incorporate any minor changes (with proper evaluation).

6. **Procedure & Training:**
    - SOP for plate counting now includes: "Plates are scanned by XYZ Automated Colony Counter. The system provides a count and identification. The analyst reviews the annotated image and result. If the image is unclear or the count is suspect (e.g., colonies very crowded or touching), the analyst shall manually count or recolonize as needed." Essentially, instruct them not to just accept bad output. Also, "If the system reports 'TNTC' or flags uncertainty, handle per manual procedure..."
    - Also include a step to periodically verify the system: maybe once a week, run a control plate or have an analyst double-count one plate to ensure it's still working.
    - Analysts get training on how to use the machine and software, how to interpret the output (e.g., the system highlights each colony with a dot – they should quickly scan if any colonies un-highlighted or mis-highlighted and if so, correct).
    - Maintenance: If the AI model is static, we treat it like an instrument calibration – e.g., maybe do a yearly challenge with known plates to confirm it hasn't drifted (since static it shouldn't, but camera calibration could drift).
    - If the model can be updated (e.g., vendor releases improved model), that's a change control with re-validation of at least the OQ tests. If our own data is used to refine it, that is a whole retraining effort to treat carefully.

Documentation:

- We'd have an overall Validation Plan for the system.
- IQ protocol, OQ/PQ protocols with test cases covering counting accuracy, identification accuracy, integration, user functions, error handling.
- We might attach raw data of counts and differences.
- Risk assessment doc showing how we addressed each risk (like near-spec plates flagged, etc.).
- Vendor documentation (user manual, any algorithm description they can share, maybe their IQ/OQ if they provided one, and we complement with our specific tests).
- Final report, approving it for use.
- Traceability matrix mapping URS to test cases (always good to show everything was tested).

- SOPs, forms updated (e.g., maybe the test record now prints "Result from auto counter").
- If there's a regulatory submission aspect (like if this was considered a new method needing regulatory notification), that's outside our scope here, but something QA/reg affairs would handle.

Through this example, we see:

- **For lab automation with AI,** treat it similarly to validating an analytical instrument:
  - Define accuracy requirements.
  - Use known reference samples for validation.
  - Ensure users can verify and override results (since human still responsible for results in GMP lab).
  - Data integrity is key – images and results must be attributable and reviewable.
- **The AI part (image analysis)** is validated by comparing to the gold standard (human counts). We needed to incorporate acceptance criteria scientifically (like allowed variance).
- **Continuing control** might mirror how we do system suitability or calibration checks in labs. Perhaps include a daily check: maybe a plate with fixed dots printed can be run each day to see if count is exactly known number – like a calibration verification.
- Also, because this directly affects quality data, regulators will be keen on its validation. We might cite that *Annex 11 and 15 require validated computerized systems and methods; Annex 22 adds that AI models must be proven for intended use with test data quality metrics (*picscheme.org) – which we did by formal testing.

General Tips and Good Practices

Whether it's decision support or lab automation (or any AI application), here are some general tips I've distilled:

- **Involve End-Users Early:** The people who will use the system (be it QA staff or lab analysts) should be part of validation planning and risk assessment. They will tell you practical failure modes you didn't think of, and having their buy-in means smoother adoption. Plus, you can design user interfaces or procedural steps that fit their workflow – which is part of making the system work as intended.
- **Leverage Simulation and Sandboxes:** Especially for decision support, it's helpful to simulate scenarios during testing. E.g., for the deviation assistant, simulate a backlog of deviations and see how the tool prioritizes them – does it align with what a senior QA person would do? This qualitative evaluation can be part of PQ or UAT (User Acceptance Testing).

- **Don't Skip Negative Testing:** Try to break the AI or confuse it in testing – feed a completely nonsense input, or a worst-quality input, and see how it reacts. The system should handle it gracefully (error message or flag) rather than silently giving a wrong output. For example, put a hair on a colony plate to see if the system mistakes it for a colony – if yes, maybe add a detection for foreign objects or ensure review.
- **Document the Rationale for Any Acceptance of Imperfection:** AI usually won't be perfect. So you might accept some error rate. It's important to justify *why that residual error risk is acceptable*. For instance: "The colony counter may undercount by up to 5%. This is deemed acceptable because specifications have a built-in safety margin and any undercount is unlikely to mask an out-of-spec result; plus, analysts will visually spot large undercounts (more than a few colonies difference)." This kind of rationale shows you thought about the impact.
- **Use Templates for Efficiency:** For AI systems, I created some document templates that help ensure we cover everything:
    - *AI System URS Template:* includes sections for data requirements and performance metrics (which typical URS might not mention).
    - *Model Evaluation Protocol:* a template to test an AI model's predictive performance – basically a structured way to do what data scientists might do informally. It ensures we capture outputs, calculate metrics, and have predefined acceptance.
    - *Monitoring Log Template:* e.g., a simple form or Excel where we periodically record key performance indicators (like monthly accuracy on new cases) as part of ongoing verification. Refer to Appendix B for more.
    - *Change Control Checklist for AI:* reminding to consider retraining data, whether old vs new model have equivalent performance, etc., when a change is proposed.
- **Be Ready to Explain to Inspectors:** During an audit, I had an inspector ask: "How do you know this fancy AI isn't making mistakes that you're not catching?" I walked them through our validation and controls – particularly how we do ongoing monitoring and how humans oversee. It satisfied them, but it reminded me that we should always frame our validation reports in a way that answers that very question. Highlight the risk-based testing approach in the validation summary: e.g., "We identified the highest risk scenario as X, and tested it by Y, confirming Z." This makes it clear we proactively tackled the critical points.
- **Stay Current:** AI tech and guidances evolve. Subscribe to updates from regulators or industry forums. For example, FDA might release new guidance on AI in manufacturing or ICH might come out with something. Showing awareness of the latest (like referencing the 2025 FDA draft guidance or the PIC/S Annex 22 even if it's draft) indicates you're aligning with current thinking, which auditors appreciate.

In conclusion for this section, validating AI in decision support or lab automation is entirely achievable with our existing validation toolbox – we just extend it to cover data and model performance. The framework is: define precisely what the AI should do, ensure it does that through testing with real examples, provide a safety net (usually human oversight or additional checks) for when it errs, and document everything. With that done, you can confidently deploy these advanced systems and enjoy their benefits (faster deviation reviews, less tedious colony counting, etc.) while staying compliant and audit-ready.

Now, as we approach the end of the guide, I'll provide some final resources: templates you might find useful, a glossary of terms we've used (for quick reference), and the full list of references to regulatory documents and sources I cited. These will help you turn this guidance into action in your validation projects.

7. Templates, Glossary, and References

In this final part, I've compiled some additional resources to help you implement what we've discussed:

Templates

(Below are outlines of templates/documents that can be helpful when validating AI systems. You can adapt these to your organization's format.)

- **AI Validation Plan Template:** A plan that outlines scope, team, milestones, and references applicable guidances. Include sections for *System Description*, *Risk Assessment Summary*, *Validation Strategy* (e.g., which lifecycle phases, what testing will be done for the model vs. the overall system), and *Acceptance Criteria*. Make sure to mention compliance with relevant guidelines (e.g., "This plan aligns with FDA 2025 AI guidance and Annex 22 requirements for model validation (2025, FDA Draft Guidance; 2025, EMA Draft Annex 22)").
- **Requirements Specification Template (with AI extensions):** Start with your typical URS/FRS format, but add sections such as:
    - *Data Requirements:* (e.g., "Training data shall include at least 3 years of historical batch data covering seasons").
    - *Performance Requirements:* (quantitative metrics for the AI model outputs).
    - *Regulatory/Compliance Requirements:* (e.g., audit trail, security roles for AI system).
    - *Scope Limitations:* explicitly state what the system will not do or handle (this helps define context of use as regulators suggest - refer to appendix c).
- **Risk Assessment Worksheet:** A table listing identified risks, their analysis (severity/probability/detectability or a risk score), existing controls, additional controls needed, and final residual risk. Include AI-specific rows (like "Model misclassification of X leads

to Y consequence – High severity – Control: procedure requires human confirmation, etc."). This can follow FMEA style. Remember to reference QRM principles (ICH Q9) for terminology, since inspectors know that language.

- **Test Plan/Protocol for AI Model Performance:** A template that enumerates test scenarios for the AI's predictive accuracy. For each test scenario:
    - Objective (e.g., "Verify model correctly classifies defect vs. good images at required accuracy"),
    - Test Data to be used (with reference to a prepared dataset),
    - Expected outcome/acceptance (e.g., "≥95% classification accuracy on the set" or "no critical false negatives observed").
    - Actual outcome. Include also procedural tests (like integration or UI) in the same or separate protocol.
- **Monitoring Log Template:** A simple document or spreadsheet that you can use to periodically record performance metrics in production (e.g., monthly accuracy, drift indicators). It might have columns for date, metric1, metric2, any anomalies, actions taken. Having a template encourages the habit of ongoing verification. Refer to Appendix B for electronic examples.
- **Change Control Impact Assessment Form (AI-specific):** This form can be used when proposing changes to an AI system (like retraining model, changing algorithm, expanding scope). It guides the user to consider:
    - Does the change affect the model's intended use or context?
    - Does it require new training data or produce a new model version?
    - What verification will be needed (perhaps re-run certain test cases, or full re-validation)?
    - Regulatory notification needed? (Probably not for internal systems, but if it's part of an approved process, consider if any regulatory filing is impacted.)
    - Who needs to approve? (Include QA and possibly data science expert).
      This ensures changes are controlled as per Annex 22's expectations on change control.

These templates are starting points – they must be integrated with your company's quality system to be effective. The key is to explicitly address data, model performance, and continuous control in your documentation, which traditional templates might not have covered.

8. Glossary

Adequacy (of AI model): the model performs reliably, consistently, and safely within its defined operating domain for the intended use, with all known risks reduced to an acceptable level.

AI (Artificial Intelligence): In this context, a broad term for computer systems able to perform tasks that typically require human intelligence. Examples include decision-making, visual perception, and language understanding. Often achieved through machine learning or rule-based logic.

ML (Machine Learning): A subset of AI focused on algorithms that improve their performance as they are exposed to more data. Instead of being explicitly programmed for every scenario, ML models learn patterns from training data. *Supervised learning* involves learning from labeled examples, *unsupervised learning* finds patterns without explicit labels, and *reinforcement learning* learns via feedback/rewards.

CSV (Computer System Validation): A systematic approach to prove (with documented evidence) that a computerized system does what it is intended to do in a consistent and reliable manner, in alignment with GxP regulations. Emphasizes verifying system fitness for use, data integrity, and compliance with requirements like 21 CFR Part 11 (for electronic records/signatures).

GMP (Good Manufacturing Practice): Regulations requiring manufacturers to ensure products are consistently produced and controlled to quality standards. Relevant here because any system used in GMP processes (including AI systems that influence manufacturing or QC decisions) must comply with these principles (e.g., proper documentation, change control, traceability).

GxP: General term for "good practice" quality guidelines and regulations in life sciences (the "x" stands for manufacturing (GMP), laboratory (GLP), clinical (GCP), etc.). An AI system might touch multiple GxP areas (e.g., GCP if used in trials, but our focus is GMP/GLP environment).

FDA Draft Guidance (2025): Refers to *"Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products"* – a January 2025 draft guidance by FDA. It provides a framework to assess AI models' credibility for supporting decisions, emphasizing context of use, risk assessment, and lifecycle management of models (FDA, 2025).

EMA Reflection Paper (2024): The *"Reflection paper on the use of AI in the medicinal product lifecycle"* issued by EMA in 2024. It outlines principles and points to consider when applying AI/ML in any stage of drug development, manufacturing, or post-market, highlighting the need for human-centric design, risk management, bias control, etc. (EMA, 2024).

Annex 11: Part of EU GMP guidelines, covering computerized systems. It sets general expectations for system validation, data integrity, security, etc. When adding AI, Annex 11 still applies plus the new Annex 22 specifics.

Annex 22 (Draft): A proposed new annex to EU/PIC/S GMP specifically for AI and ML. It introduces requirements for using AI in manufacturing, such as defining intended use, ensuring data and model quality, validation of models, and continuous oversight (EMA, 2025 draft).

PIC/S: A global cooperative forum of pharmaceutical inspectors. PIC/S adopts many EU GMP guidelines. The PIC/S Annex 22 is essentially the same guidance as EMA's Annex 22, ensuring global alignment.

ISPE GAMP 5 Second Edition: Industry guide (2022) by International Society for Pharmaceutical Engineering. GAMP 5 provides a risk-based approach to validating computerized systems. The second edition updated guidance for modern tech like AI/ML (Appendix D11) and encourages critical thinking (Computer Software Assurance) instead of a documentation-heavy approach, while still meeting regulatory needs.

Context of Use (COU): A term often used in FDA docs, referring to the specific conditions and purpose for which an AI model is intended. Defining COU means describing where, how, and by whom the model will be used, and what decision it supports. Validation is done against that specific context (appendix C).

Credibility (of AI model): In regulatory terms, the trustworthiness of an AI's output for its intended use. Establishing credibility involves demonstrating the model is scientifically valid, accurate, and reliable enough for the decision at hand (through tests, evidence, etc.). FDA's framework focuses on credibility assessment (appendix D).

Quality Risk Management (QRM): A systematic process for assessing, controlling, communicating and reviewing risks to quality (ICH Q9 guideline). We use it to manage AI-related risks in GMP, by identifying potential failures and ensuring controls are in place, commensurate with risk.

Bias (AI Bias): Systematic error in AI outputs due to skewed training data or algorithm design. For example, a model might consistently under-predict values for a certain subgroup because that subgroup was underrepresented or different in the training data. Managing bias involves using diverse data and testing the model for unfair/systematic errors.

Overfitting: When an ML model is too closely fitted to the training data, capturing noise instead of the underlying pattern. Overfit models perform well on training data but poorly on unseen data. Detection involves comparing performance on training vs. test sets (refer to appendix E for bias and overfitting details).

Drift (Data/Concept Drift): Over time, the statistical properties of input data (or the relationship between inputs and outputs) may change, causing model performance to degrade. *Data drift* means the input distribution changes (e.g., new raw materials cause different sensor readings), *concept drift* means the actual underlying relationship changes (e.g., a process equipment's behavior changes as it ages). Monitoring drift is essential to know when retraining might be needed (appendix F).

Human-in-the-Loop (HITL): A design where human oversight is retained in an AI-assisted process. The AI provides output or suggestions, but a human reviews or must approve before final decisions. HITL is a key risk mitigation approach in GMP for high-stakes uses of AI (appendix G).

Explainability: The extent to which an AI system's workings and outputs can be understood by humans. High explainability means you can trace why the model gave a certain output (e.g., which factors were important). Regulators encourage using interpretable models or providing explanations, especially for decisions impacting quality or safety.

Validation (of AI model vs. system): We often talk about validating an AI model's performance (does it meet accuracy criteria, etc.) *and* validating the overall system it's integrated in (does the software/hardware around it function correctly in the process). Both are needed: model validation is like a subset focusing on algorithm output, system validation covers end-to-end functionality.

Performance Metrics: Quantitative measures of model output quality. Examples: accuracy, precision, recall (sensitivity), specificity, error rate, etc. Chosen based on the use case (e.g., for anomaly detection, might use false positive/negative rates; for regression predictions, use mean error). Setting acceptance criteria on these is part of validation.

Training Set / Test Set: Subsets of data used in machine learning. The training set trains the model, and the test set is a separate data used to evaluate the model's performance objectively. Sometimes a *validation set* is also used in development to fine-tune parameters. Annex 22 requires keeping test data independent of training to ensure a valid performance assessment.

Change Control (for AI): The process of managing changes to the AI system in a controlled manner. This could mean changes to the model (retraining, new algorithm version) or the data pipeline or parameters. Any such change should go through formal approval with assessment of whether re-validation is needed (spoiler: usually yes, at least in part). The FDA PCCP concept and Annex 22 both emphasize predefined plans for managing model updates (health.ec.europa.eu).

21 CFR Part 11: U.S. regulation on electronic records and signatures. If your AI system records data electronically that is used to make quality decisions, ensure compliance (secure,

timestamped, audit trails, etc.). Similarly, EU Annex 11 addresses these principles. While not AI-specific, these regs still apply to the systems we're validating.

9. References

(Below are the full references for sources and regulations cited in this guide, with URLs for accessibility - the closest URL is provided if source is behind a paywall.)

- Association for the Advancement of Medical Instrumentation. (2023). *AAMI TIR34971:2023 – Application of ISO 14971 to machine learning in artificial intelligence – Guide*. Arlington, VA: AAMI. *https://webstore.ansi.org/standards/aami/aamitir349712023*

- European Medicines Agency (EMA). (2024). *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle*. EMA/CHMP/CVMP/83833/2023. Retrieved from https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle

- European Medicines Agency (EMA). (2025). *Draft Annex 22: Artificial Intelligence (Consultation Draft).* In EudraLex Volume 4: EU Guidelines for GMP (Annex 22). Retrieved from https://health.ec.europa.eu/document/download/annex22_artificial-intelligence_draft.pdf (Public consultation draft issued July 2025)

- European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union, L 259, 1–81. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

- Food and Drug Administration (FDA). (2025). *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products* (Draft Guidance for Industry). FDA Docket FDA-2024-D-4689. Retrieved from *https://www.fda.gov/media/184830/download*

- Pharmaceutical Inspection Co-operation Scheme (PIC/S). (2025, July 7). *Joint stakeholders consultation on the revision of Annex 11 and new Annex 22 on Artificial Intelligence of the PIC/S and EU GMP Guides* (News article). PIC/S Official Website. Retrieved from https://picscheme.org/en/news/joint-stakeholders-consultation-on-the-revision-of-chapter-4

- Health Canada. (2019). *Guidance document: Software as a medical device (SaMD) – Definition and classification*. Ottawa, ON: Health Canada. *http://publications.gc.ca/pub?id=9.882300&sl=1*

- International Organization for Standardization. (2020). *ISO/IEC TR 29119-11:2020 – Software and systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems*. Geneva, Switzerland: ISO. *https://www.iso.org/standard/79016.html*

- International Organization for Standardization. (2021). *ISO/IEC TR 24027:2021 – Information technology – Artificial intelligence (AI) – Bias in AI systems and AI-aided decision making*. Geneva, Switzerland: ISO. *https://www.iso.org/standard/77607.html*

- International Organization for Standardization. (2022a). *ISO/IEC 22989:2022 – Information technology – Artificial intelligence – Artificial intelligence concepts and terminology*. Geneva, Switzerland: ISO. *https://www.iso.org/standard/74296.html*

- International Organization for Standardization. (2022b). *ISO/IEC 23053:2022 – Framework for Artificial Intelligence (AI) systems using Machine Learning (ML)*. Geneva, Switzerland: ISO. https://www.iso.org/standard/74438.html

- International Organization for Standardization. (2023). *ISO/IEC 42001:2023 – Information technology – Artificial intelligence – Management system*. Geneva, Switzerland: ISO. *https://www.iso.org/standard/81230.html*

- International Society for Pharmaceutical Engineering. (2025). *GAMP® Guide: Artificial Intelligence*. North Bethesda, MD: ISPE. https://ispe.org/publications/guidance-documents/gamp-guide-artificial-intelligence

- International Society for Pharmaceutical Engineering (ISPE). (2022). *GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems* (Second Edition). ISPE Guidance Document. (Summary of updates retrieved from CAI blog: https://caiready.com/life-sciences/blog/use-of-artificial-intelligence-and-machine-learning/)

- Rephine. (2025, August 15). *How to Prepare for Annex 22: EMA's AI and Machine Learning Guidance for Pharma Manufacturing*. [Blog post]. Retrieved from https://www.rephine.com/resources/blog/how-to-prepare-for-annex-22-emas-ai-and-machine-learning-guidance-for-pharma-manufacturing/

- NSF International. (2024, September 30). *EMA Reflection Paper on AI in the Medicinal Product Lifecycle*[Regulatory news]. Retrieved from https://www.nsf.org/ca/en/ life-science-regulatory-news/ema-reflection-paper-on-ai-medicinal-product-lifecycle

(The above references provide additional context and are the sources of the inline citations used in this guide. When implementing AI validation, always refer to the latest official guidelines and documents from regulatory agencies for the most up-to-date information.)

Closing note: Validating AI systems in a GMP world might seem complex, but it boils down to demonstrating control over the system's behavior and assuring quality – concepts we live by in all of GMP. By understanding the technology, applying sound risk management, and following regulatory guidance, we can confidently integrate AI innovations into our processes. I hope this guide has demystified the process and given you a clear roadmap to follow. Good luck with your AI validation projects, and welcome to the evolving future of CSV in the age of AI!

10. Appendices

APPENDIX A: AI MODEL ADEQUACY

Subpart 1: What "Good Enough" Means for AI Models in GMP

There is **no universal accuracy target,** no required percentage, and no single metric that regulators mandate. Instead, "good enough" means:

The model performs reliably, consistently, and safely within its defined operating domain for the intended use, with all known risks reduced to an acceptable level.

This is *always* decided based on:

(a) The severity of harm if the model is wrong
(b) The role of human oversight
(c) The criticality of the decision
(d) The strength of the validation evidence

How "Good Enough" Is Normally Determined

1. Use Case Criticality

The more important the decision, the higher the performance requirement.

Examples:

- **Low criticality:** Suggesting keywords in a batch record. A model might be acceptable at 85 percent accuracy.
- **Medium criticality:** Sorting lab data or flagging anomalies. Often requires 90 to 95 percent performance.
- **High criticality:** A decision support tool that influences a batch disposition
  - Requires near deterministic reliability
  - Strong guardrails and human verification
  - Often 98 percent or better depending on the failure modes. This is not about "the FDA says 98 percent", it is about: What level of residual risk is acceptable for the GMP process this model touches.

2. Acceptable Error Types

- Accuracy alone does not define adequacy.
- You evaluate both error *rate* and error *type*.

Examples:

- A model that occasionally *misses* minor anomalies may be acceptable.
- A model that *creates false anomalies* 5 percent of the time may not be acceptable because it disrupts operations.
- In GMP, false negatives and false positives have **different consequences**, and your threshold depends on which matters more.

3. **Acceptance Criteria Defined Before Validation**

   Regulators expect you to set these *before* you test.

   Acceptance criteria usually include:

   - Minimum performance metric thresholds
   - Boundary tests for edge cases
   - Stability across multiple datasets
   - Consistency across repeated runs
   - No catastrophic errors
   - No unsafe or misleading outputs

   The model is "good enough" only if it **meets all defined criteria.**

4. **Benchmarking Against a Human or Legacy System**

   A common method in the field:

   Model performance must meet or exceed the existing process baseline.

   Examples:

   - If human reviewers catch 92 percent of anomalies, then a model must be at least 92 percent or higher.
   - If a legacy system misroutes 4 percent of lab samples, the AI must perform better than that.

   This is often the strongest, easiest, and most defensible argument in an audit.

5. **Residual Risk Evaluation**

   Even if metrics are strong, you must explicitly answer:

   What risks remain, and are they acceptable given the controls and human oversight?

   If residual risk is unacceptable:

   - The model is *not* "good enough"
   - You must retrain, adjust, narrow the operating domain, or redesign guardrails

6. **Human Oversight Level**

   Human review changes what "good enough" means.

   With strong human review: A model might be acceptable at:

   - 85 percent
   - 90 percent
   - 93 percent

With weak or no human review:

You need:

- very high performance
- robust guardrails
- deterministic fallback behavior

Is There a Percentage Goal?

No fixed number exists. But in practice, across regulated environments:

- **80 to 90 percent** is common for low criticality tasks
- **90 to 95 percent** for medium
- **95 to 99 percent** for high
- **99 percent+** only when failure risks are very high, and the model influences product quality or patient safety

These ranges *are not mandated* but reflect patterns from risk assessments and industry expectations.

7. "Good Enough" = All of These Together

A model is adequate only when:

1. It meets predefined acceptance criteria
2. It performs well within its operating domain
3. Its risks are mitigated
4. Human oversight is appropriate
5. It behaves consistently
6. It *reduces* risk compared to the current method
7. Residual risk is acceptable

If any of these fail, the model is **not** "good enough."

Subpart 2: Adequacy "Good Enough" Decision Tree (for GMP)

A practical framework for determining if a model is adequate for intended use.

1. Start: Did the model meet predefined acceptance criteria?

If you did not define acceptance criteria before testing, stop and define them.

Yes → Go to Step 2
No → Not good enough. Retrain, adjust, or redefine scope.

———

2. Is the model's performance stable across all test sets? This includes:

- primary validation set
- secondary or stress sets
- edge cases
- multilingual sets if relevant
- low frequency but high impact scenarios

Stable → Go to Step 3
Unstable → Not good enough. Investigate variability or narrow operating domain.

———

3. Are the errors acceptable for the risk level of the use case? Look at:

- false positives
- false negatives
- error type severity
- error frequency
- business impact
- GMP risk impact

Errors acceptable → Go to Step 4
Errors unacceptable → Not good enough unless you add controls or oversight.

———

4. Are all critical risks mitigated with controls or guardrails?

Examples:

- human confirmation
- restricted actions
- forced justification
- RAG grounding
- blocked unsafe content
- alerts for low confidence

All mitigated → Go to Step 5
Not mitigated → Add controls or redesign.

———

5. Can the model stay inside its approved operating domain? Check:

- training data coverage
- drift susceptibility
- prompt sensitivity
- hallucination rates
- domain boundaries
- confidence signals

Yes → Go to Step 6
No → Reduce domain or strengthen guardrails.

———

6. Does the model perform at least as well as the existing process baseline?

Examples of baselines:

- human SME accuracy
- manual review rate
- legacy algorithm performance
- historical false negative rate

Meets or exceeds → Go to Step 7
Below baseline → Not good enough. Improve or retrain.

———

7. Is the planned level of human oversight appropriate for the residual risk?

If residual risk is high, oversight must be strong.
If residual risk is low, oversight can be lighter.

Oversight appropriate → Go to Step 8
Oversight inadequate → Add human review or reject the model.

———

8. Are residual risks acceptable when viewed together?

This is a final holistic check:

- performance
- consistency
- risk profile
- operating domain
- oversight
- guardrails

Residual risks acceptable → Model is good enough
Residual risks unacceptable → Do not deploy

Decision Tree Summary

1. Meets acceptance criteria?
2. Stable across test sets?
3. Errors acceptable for risk?
4. Critical risks mitigated?
5. Stays inside operating domain?
6. Meets or exceeds baseline?
7. Oversight appropriate?
8. Residual risk acceptable?

If **any** answer is no, the model is **not good enough** until addressed.

Subpart 3: Sliding scale for AI Adequacy, "Good Enough" in GMP

The higher the criticality, the higher the bar. Performance expectations scale with the impact of failure.

1. Low Criticality

Supportive tasks with no direct impact on product quality or patient safety

Examples:

- Suggesting keywords
- Drafting controlled content
- Grouping documents
- Non binding alerts

Performance Expectation:

80 to 90 percent accuracy (or comparable metric)

Tolerable Error Types:

- Occasional misses
- Low consequence false positives

Oversight Level:

Light human review or sample checks

Adequacy Threshold:

"Good enough" if it performs **as well as or better than the current manual process.**

SLIDING SCALE: AI "GOOD ENOUGH" IN GMP

Low Criticality	Medium Criticality	High Criticality
80–90% Accuracy	**90–95%** Accuracy	**95–99%** Accuracy
Occasional misses	Very few false negatives	None that cause risk
Light ove-rsight	Moderate oversight	Strong oversight
Supports tasks	Influences decisions	Impacts quality decisions

2. Medium Criticality

Tasks that influence decisions but do not make final decisions

Examples:

- Lab anomaly detection suggestions
- Pre sorting QC data
- Summarizing batch deviations
- Automated extraction of instrument parameters

Performance Expectation:

90 to 95 percent accuracy (or relevant metric), stable across primary and secondary datasets

Tolerable Error Types:

- Very few false negatives
- Few but understandable false positives

Oversight Level:

Regular human confirmation of outputs

Adequacy Threshold: "Good enough" if:

- model meets all acceptance criteria
- errors are explainable and controlled
- risk controls reduce residual risk to acceptable levels

3. High Criticality

AI influences product quality decisions, batch status, or safety related conclusions

Examples:

- Decision support for batch impact assessments
- OOS triage logic
- Root cause suggestion tools
- Safety related extraction from lab systems

Performance Expectation:

95 to 99 percent+ accuracy, near deterministic behavior, highly consistent across stress tests

Tolerable Error Types:

- None that cause product quality or safety risk
- False negatives must be extremely rare

Oversight Level:

- Strong and documented human review
- Final decision always made by an experienced person

Adequacy Threshold: "Good enough" only when:

- performance exceeds the existing human or system baseline
- all high risk failure modes are mitigated
- the model behaves predictably
- residual risk is demonstrably low

4. Very High Criticality (Rare in GMP AI)

AI used in manufacturing control or real time automated actions

(Most companies avoid this category entirely)

Performance Expectation: Approaches deterministic, almost no tolerance for incorrect outputs

Adequacy Threshold: Requires extensive justification and restrictions thus most organizations decide not to deploy AI here

APPENDIX B: AI MONITORING PRACTICES

Almost nobody uses paper logs for AI monitoring in regulated environments. Even in GMP, where paper is still used for many things, **AI monitoring requires modern tooling**, because the data volumes, drift signals, and model behaviors cannot be reliably captured on paper.

Here's what organizations actually use in practice:

———

What Companies Really Use for AI Monitoring (GMP-appropriate)

1. Modern Digital Dashboards (Most Common)

Platforms like:

- Azure ML Monitoring
- Amazon SageMaker Model Monitor
- Datadog ML Observability
- Fiddler AI
- EvidentlyAI
- Weights & Biases

These systems create:

- continuous data drift charts
- output distributions
- accuracy over time
- alerts for unusual patterns
- confidence spread graphs
- feature drift heat-maps

Why GMP teams like it:
You can export PDF snapshots monthly into your QMS to satisfy compliance.

———

2. CSV-Ready Monthly Summary Reports (Very Common in Pharma)

These are structured documents generated automatically, but formatted in a validation-friendly way:

- monthly PDF summary
- tables of metrics
- graphs and drift indicators
- explainability changes
- human review signatures

Why regulators like it:
It mirrors batch review — predictable and controlled.

———

3. Electronic Logs (Excel or Controlled Spreadsheet in QMS)

Mid-maturity organizations often use:

- **controlled Excel** with locked formulas
- **versioned SharePoint sheets**
- **logs inside Veeva or MasterControl**

This is essentially your monitoring log, but:

- increased legibility
- version controlled
- validated
- audit trailed

———

4. Automated Alerts Integrated with QMS

Some organizations integrate monitoring alerts into:

- **TrackWise**
- **MasterControl eQMS**
- **ZenQMS**
- **Veeva Vault**

For example:

- If drift exceeds 0.10 → auto-generate an "AI Monitoring Event"
- If accuracy drops below threshold → create a CAPA record

This creates a **GMP-aligned, audit-friendly trail.**

———

5. Validation-Bound Monitoring Notebooks (for data science teams)

Teams often maintain:

- Python notebooks
- R scripts
- scheduled dashboards
- Jupyter pipelines

These produce visualizations and numerical summaries. Outputs are exported into QMS monthly. **These are not GMP records**, but they feed the official GMP log.

———

6. Paper Logs

Paper is only used when:

- the AI system is extremely simple
- the use case is low criticality
- monitoring is based on small, discrete checks

Paper is troublesome in handling:

- drift metrics
- distribution plots
- confidence spreads

- statistical monitoring

I recommend using paper logs to finalize your workflows then converting to digital, if your AI system doesn't have monitoring built in.

What Inspectors Expect

Inspectors expect a **quality-managed digital record**, not literal paper. They want to see:

- evidence of monthly review
- thresholds and actions
- drift detection
- performance tracking
- signatures and dating
- traceability to CAPA where needed

A clean monitored **digital table or PDF summary** fully satisfies regulators, especially if integrated into your QMS.

APPENDIX C: CONTEXT OF USE

Context of Use (COU): What It Really Means in FDA/EMA Language

"Context of Use" (COU) is one of the most important concepts in validating AI systems. Regulators use it to anchor what the model is allowed to do, what it is not allowed to do, and how validation must be scoped. In FDA and EMA documents, COU generally includes:

1. The intended purpose

What problem the AI solves. Example: "Automatically classify QC instrument errors to aid lab troubleshooting."

2. The decision or action the AI influences

What the user will actually do with the output. Example: "Assist analysts in determining whether an OOS is due to instrument failure or sample issue."

3. Who uses it

Their training level and role. Example: "QC analyst I and II, supervised by QC lead."

4. When and where it is used

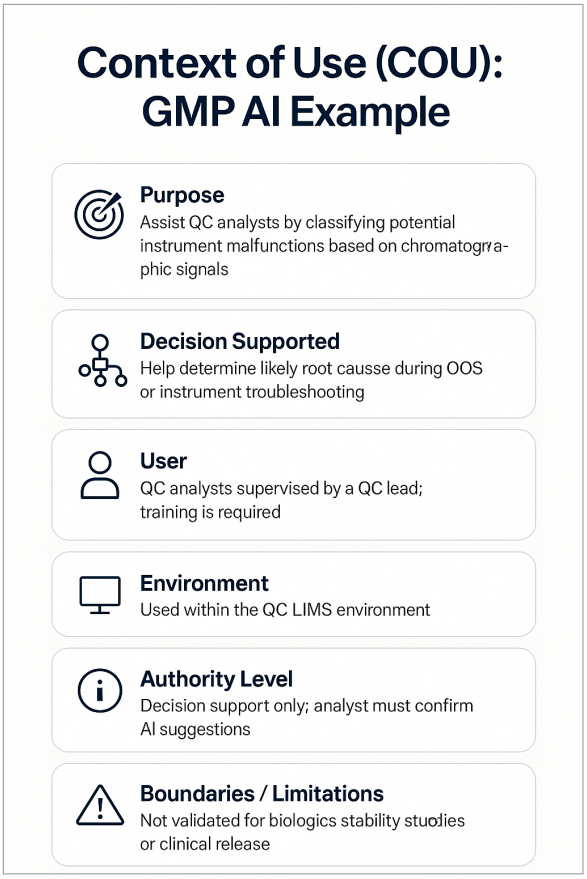
Environment and business process step. Example: "Used during initial triage of instrument anomalies in the QC lab."

5. How much authority the model has

Supportive suggestion vs automated decision. This defines the criticality and the required performance.

6. Boundaries and limitations

What the model must not be used for. This protects from misuse (a major regulatory concern).



Why COU Matters in Validation

Regulators expect the **validation to match the COU,** not a generic standard. This means:

- Test data must match the scenarios described in the COU
- Risk assessment must reflect the decisions the model influences
- Acceptance criteria must correspond to the impact of those decisions
- Monitoring must track metrics relevant to that specific use
- Documentation must show that the model works **in that exact context**

If the COU changes, the validation must change with it.

GMP Case Study: COU for an AI Model in QC Lab

Scenario Overview

A biotech company implements an AI model that classifies QC chromatography instrument failures based on patterns in system suitability tests and injection characteristics.

The model does **not** make final decisions. It provides **decision support** to reduce investigation time.

———

1. Context of Use Statement (Realistic Example)

Purpose:
Assist QC analysts by classifying potential instrument malfunction types (e.g., detector drift, pump pulsation, sample contamination) based on raw chromatographic signals.

Decision Supported:
The system provides **suggested classifications** to help analysts determine likely root cause during the initial stages of an OOS or instrument troubleshooting.

User:
QC analysts (Level I–III), supervised by a QC lead. Users must complete training on AI behavior, limitations, and oversight requirements.

Environment:
Used within the QC LIMS environment during routine testing and OOS investigations.

Authority Level:
Decision support only. AI suggestions must be reviewed and confirmed by a qualified analyst. The AI output **cannot be used to close an OOS** or to make a final quality decision.

Boundaries / Limitations:

- Not validated for biologics stability studies
- Not validated for impurity profile analysis
- Not validated for clinical release lots
- Must not be used when instrument baseline noise exceeds predefined thresholds
- Not certified for final decision making

———

2. How COU Drives the Validation Strategy

Test Data Must Match the COU

Since the model is used for chromatography troubleshooting:

- Validation dataset includes chromatograms from at least 2–3 years
- Includes all instrument failure modes documented in deviation logs
- Includes edge cases (partial pump failures, low-signal drift)

Performance Requirements Must Reflect Decision Support

Because the model is **decision support**, not final action:

- Accuracy threshold: **90–93 percent**
- False negative tolerance: low (high severity)
- False positive tolerance: slightly higher (medium severity)

Risk Assessment Uses COU to Define Criticality

Failure mode: "Model incorrectly classifies detector drift as sample contamination."
Impact: Delays investigation but does **not** cause incorrect batch disposition because analyst reviews it.
Criticality: Medium.

Oversight Aligned to COU

Human confirmation step required:

- Analyst must review AI classification
- Analyst must acknowledge when model output is overridden
- System must display "AI-generated suggestion — not a final decision"

Monitoring Aligned to COU

Monthly drift monitoring includes:

- Accuracy per failure mode
- Distribution shift in chromatographic inputs
- Confidence spread on low-signal runs
- Instances where analysts override the AI suggestion

———

3. Case Study Outcome

After the validation:

- System met acceptance criteria
- Residual risk acceptable due to human oversight
- Monitoring plan approved by QA
- COU clearly limits use to triage decision support
- SOP updated to reflect COU boundaries
- Analysts trained on oversight, limitations, and confirming AI suggestions

The model was approved for controlled deployment.

———

4. Why This Case Study Works

This example shows what regulators look for:

- Clear purpose and limits
- Clear actions the AI influences
- Clear roles and training
- Clear alignment between COU and validation evidence
- Clear monitoring expectations

This is exactly how inspectors expect AI COU documentation to look during a GMP audit.

APPENDIX D: CREDIBILITY OF AI MODELS

When regulators use the term **credibility,** they mean:

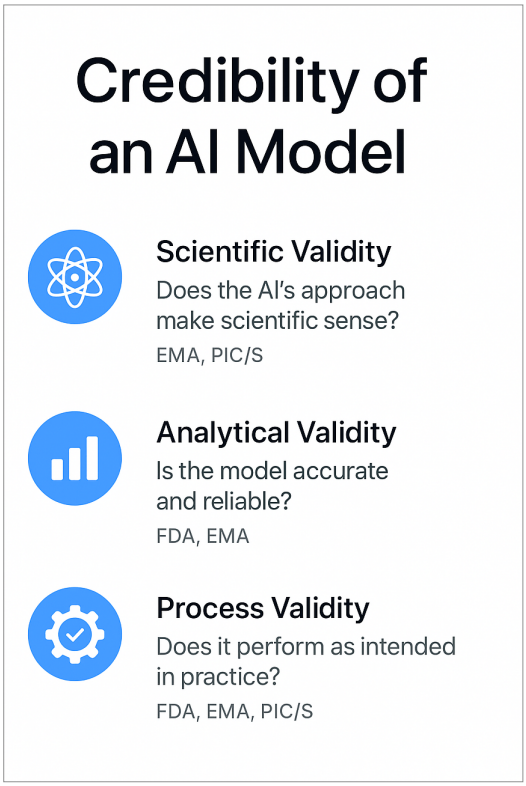
How much trust we can place in the model's output for the specific decision it supports. Credibility is not a generic quality. It is **context-dependent, evidence-based, and proportional to risk.**

A model is credible when it has enough scientific and validation evidence to support the decision it influences *in its defined Context of Use (COU).*

FDA's 2025 AI guidance, the EMA Reflection Paper, and Annex 22 all use this concept in different ways:

- FDA talks about credibility assessment
- EMA talks about scientific validity and reliability
- PIC/S talks about trustworthy performance in the approved operating domain
- ISPE talks about model evidence that aligns with risk and intended use

All point to the same thing: **credibility = evidence that the model works for its intended purpose and within its defined limits.**



The Three Components of AI Credibility. Credibility is built on three pillars that regulators explicitly reference:

---

1. Scientific Validity

Does the underlying approach make sense scientifically for the intended decision?
Examples:

- Is this algorithm appropriate for the type of data?
- Are features relevant and grounded in domain science?
- Does the model reflect known mechanisms or patterns in the process?
- Does the approach match what experts already understand?

Without scientific validity, no amount of testing can make the model credible.

---

2. Analytical Validity

Is the model mathematically and statistically sound? This includes:

- accuracy
- precision
- recall

- false negative rate
- calibration
- robustness
- repeatability
- stability across datasets
- performance in edge cases

You generate evidence through:

- holdout testing
- stress testing
- cross-site testing
- negative controls
- sensitivity analyses

This is the **bulk of model validation**.

---

3. **Clinical or Process Validity (GMP Equivalent).** This answers the question: **Does the model work reliably in the real process it is meant to support?** In GMP this means:
   - Does the model actually help the decision it claims to support?
   - Does it work with real batch records, real instruments, and real-quality workflows?
   - Do analysts find the output reliable and interpretable?
   - Does it reduce or at least not increase process risk?

   This is where *operating domain*, *human oversight*, and *risk controls* come together.

What Evidence Builds Credibility

Regulators expect the following evidence types:

✓ **Traceable requirements tied to COU**
The model must be evaluated against the exact use case.

✓ **Dataset representativeness evidence**
Training and validation data must match real work.

✓ **Performance testing across all relevant conditions**
Not just average accuracy.

✓ **Robustness testing**
Noise, drift, perturbation, RAG mis-grounding, adversarial prompts.

✓ **Bias/variability assessment**
Does the model behave differently across instruments, analysts, lots, or conditions?

✓ **Human factors evidence**
Users must be able to understand and appropriately rely on the output.

✓ **Monitoring plan**
You must prove the AI will remain credible after day one.

How Auditors Assess Credibility Quickly

Auditors will ask:

1. What decision does the model influence?
2. What evidence shows it works for that decision?
3. What conditions were tested?
4. Are the datasets traceable and representative?
5. How do you know it stays within its operating domain?
6. What happens if it is wrong?
7. Do users know how to interpret the output?
8. What ongoing monitoring exists?

This is their credibility checklist.

GMP Case Study: Establishing Credibility for an AI OOS Triage Assistant

Scenario

A biotech site deploys an AI decision support tool that classifies the likely source of OOS (out-of-specification) signals based on:

- chromatographic signatures
- analyst notes
- instrument metadata
- historical deviation patterns

The system provides **suggestions**, not final decisions.

---

1. **COU Drives the Credibility Standard**
   The AI supports OOS investigations. This is **medium–high criticality** because errors can mislead investigations but **cannot** change final batch disposition (human oversight). So credibility evidence must be strong, but not deterministic.

---

2. **Scientific Validity Evidence**
   The team documents that:
   - the selected modeling approach captures patterns known to correlate with failure modes
   - input features reflect real chromatographic physics
   - outputs map to categories already used by SMEs

   Regulators want this scientific grounding.

---

3. **Analytical Validity Evidence**
   The model is tested with:
   - 3 years of historical OOS investigation data
   - stratified sampling across product families
   - 25 edge case scenarios

- multi-site data to check transferability
- negative controls (e.g., clean data mislabeled intentionally)

Results:

- overall accuracy: 93 percent
- false negative rate: 2.7 percent
- stability across instruments: ±3 percent variation
- strong calibration curves (high confidence corresponds to higher precision)
- robust under 5 percent noise injection

This evidence builds analytical credibility.

———

4. Process Validity Evidence (GMP)

A pilot study shows:

- analysts resolve investigations ~20 percent faster
- 94 percent of AI suggestions match SME consensus
- analysts report understanding AI outputs easily
- human oversight consistently catches rare misclassifications
- no impact on batch disposition accuracy

This proves the model is fit for use *in the real process*.

———

5. Boundaries and Operating Domain

Credibility also requires proving **what the model cannot do:**

- not validated for biologics stability
- not validated for impurity profiling
- not validated for early development lots
- not used when instrument baseline noise exceeds 5 percent threshold

Clear boundaries increase credibility.

———

6. Monitoring Plan Supports Ongoing Credibility

The team monitors:

- monthly accuracy
- drift signals
- outlier rate
- confidence spread
- number of analyst overrides
- mismatch between AI classification and SME review

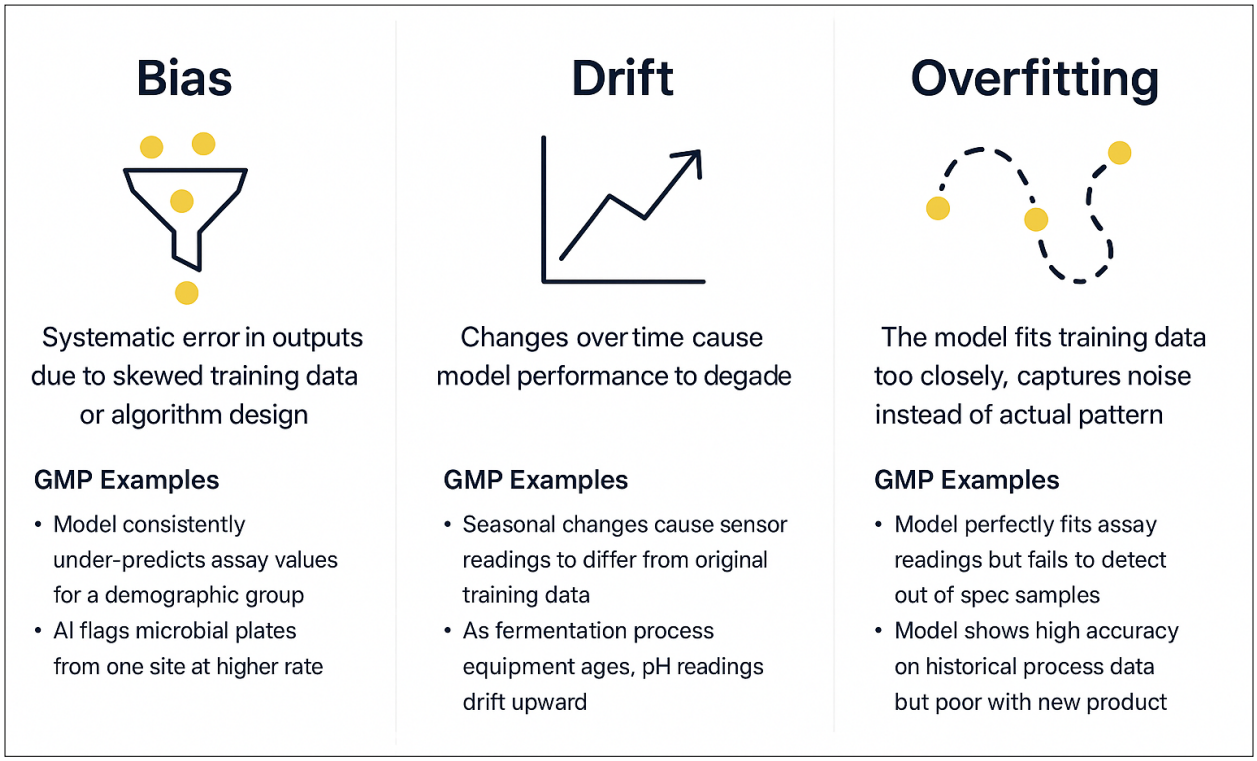
Triggers are defined:

- drift > 0.1 → generate AI Monitoring Event
- accuracy < 90 percent → CAPA

Maintaining credibility is just as important as establishing it.

———

APPENDIX E: BIAS vs OVERFITTING



Bias (AI Bias) – What It Means in Regulated AI

Bias in an AI model means **systematic, repeatable error** that affects one category of data differently than another. It does *not* mean human prejudice – it means the model behaves unevenly.

Regulators care about bias because:

- It reduces **credibility**
- It reduces **reliability**
- It violates **scientific validity**
- It causes **unseen drift**
- It can distort GMP decisions

Where Bias Comes From - Bias usually originates from 3 places:

1. **Training Data Bias**

   The training data does not represent the real population or process.

   Examples:

   - 95 percent of samples come from one facility
   - Only one model of instrument was used
   - Older lots underrepresented
   - Rare failure modes barely appear

2. Labeling / Annotation Bias

If humans label the data inconsistently, the model learns their inconsistency.
Example:
One analyst labels borderline cases differently than another analyst.

3. Algorithmic / Feature Bias

The model weights certain signals too strongly.
Example:
A model over-weights "run length" so it systematically misclassifies short-duration assays.

———

How FDA/EMA Expect Bias to Be Managed

Regulators expect:

✓ **Diverse, representative training data**
Across manufacturing lines, instruments, products, seasons, sites, lots.

✓ **Subgroup performance testing**
Does the model behave differently for:
- site 1 vs site 2
- instrument A vs instrument B
- product family X vs Y

✓ **Bias metrics**
Precision/recall differences across groups. EMA calls this "differential performance."

✓ **Documented limitations**
If the model only works for certain product families, it must be stated in COU.

✓ **Monitoring bias over time**
Bias can drift just like accuracy.

———

Case Study: Bias in QC Chromatography Failure Classification

Scenario

A model classifies instrument failure types (pump, detector, autosampler). It performs well overall (93 percent accuracy), but fails more frequently for biologics programs.

Root Cause

Training data used:
- 80 percent small-molecule programs
- 20 percent biologics

The chromatographic signatures differ across programs. The model learned a biased representation.

Detection

Subgroup testing shows:
- 95 percent accuracy for small molecules
- 82 percent accuracy for biologics

A 13 percent performance difference. Regulators classify this as bias.

Correction

- Add biologics chromatograms
- Expand failure mode library
- Re-label ambiguous examples
- Retrain model
- Re-test subgroup parity

Outcome

Accuracy levels converge:

- Small molecule: 94 percent
- Biologics: 91 percent

Bias reduced to acceptable levels.

———

Overfitting — What It Means

Overfitting occurs when the model learns **noise instead of signal.** It memorizes training examples instead of learning general rules.
This leads to:

- high training accuracy
- low test accuracy
- poor long-term reliability
- loss of credibility

Overfitting is dangerous in GMP because a model may appear excellent during validation, then collapse once real data changes even slightly.

———

How Overfitting Is Identified

1. **Performance Drop on Test/Validation Data**
   Training accuracy: 99 percent
   Test accuracy: 83 percent
   → Overfitting is almost certain.
2. **Large Confidence Spread**
   Very confident on training-like data
   Low confidence on new/rare conditions
3. **Model Behaves Unusually on Edge Cases**
   Fails under minor variations:
   - slightly different batch size
   - different detector aging profiles
   - updated instrument firmware
4. **High Model Complexity Without Justification**
   Too many parameters + not enough data.

———

How FDA/EMA Expect Overfitting To Be Controlled

- ✓ **Use separate training, validation, and test sets**
  FDA emphasizes dataset partition integrity.
- ✓ **Cross-validation**
  Run the model across multiple folds.
- ✓ **Regularization techniques**
  Dropout, weight decay, early stopping.
- ✓ **Stress and perturbation testing**
  Add synthetic noise to check robustness.
- ✓ **Representative datasets**
  Include multi-site, multi-lot, multi-instrument data.
- ✓ **Monitoring after deployment**
  If overfitting exists, drift will spike quickly.

———

Case Study: Overfitting in AI for Microbial Colony Classification

Scenario
An AI system classifies microbial colony morphology in environmental monitoring plates.

- Training accuracy = 98 percent
- Test accuracy = 78 percent
- Drift appears quickly after deployment.

Root Cause

- Model was trained on plates from only one incubator
- Lighting conditions varied in real lab
- Images were too similar in training set
- The model memorized patterns instead of learning morphology rules

Detection
During post-deployment monitoring:

- Confusion matrix shows failure to recognize rare colony types
- False negatives increase in images from a second incubator
- Confidence spread widens significantly

Overfitting confirmed.

Correction

- Expand training set across all incubators and lighting conditions
- Use data augmentation (rotation, brightness)
- Reduce model complexity
- Apply cross-validation
- Retrain and re-test

APPENDIX F: DRIFT IN GMP AI

"Drift" describes **changes over time** that cause an AI model to become less reliable.
There are **two major types**, and regulators expect you to distinguish them:

———

1. Data Drift (Covariate Drift)

Definition:
The *input* data distribution changes over time.

In GMP terms:
The signals going *into* the model no longer look like the data it was trained on.

Examples in life sciences:

- New raw material suppliers → different spectral profiles
- Updated HPLC column → different retention time patterns
- New incubator → different EM image brightness
- Seasonal changes → viscosity or temperature differences that alter sensor readings
- Upgraded firmware on a QC instrument

Why it matters:

Even if the underlying process is unchanged, the model sees **new patterns** it doesn't understand. This causes subtle but dangerous classification errors.

———

2. Concept Drift (Relationship Drift)

Definition:
The *relationship* between inputs and outputs changes.

In GMP terms:
The real-world meaning of the data changes.

Examples in life sciences:

- As equipment ages, baseline noise rises → "normal" no longer means the same thing
- A new chromatography method changes what "pump failure" looks like
- A new process change makes certain anomaly patterns benign
- A new bioreactor control loop changes the meaning of oxygen or pH spikes
- A site switches to a different lot of media → microbial growth patterns change
- Operator behavior or SOP updates change how metadata is captured

Why it matters:
A model trained before the change now **misinterprets** reality.
This can lead to poor suggestions, unnecessary investigations, or missed anomalies.

———

Why Drift Is a Big Deal for Regulators

FDA, EMA, PIC/S Annex 22, and ISPE GAMP Second Edition all emphasize:

- AI cannot be "validated once and done."
- Drift must be continuously monitored.
- Drift monitoring is required to maintain **credibility and fitness for use.**
- Drift triggers may require **retraining, model versioning, or COU limits.**

Inspectors increasingly look for drift signals **as part of your ongoing monitoring plan**.

———

How to Validate Against Drift (Practical Steps)

Validation must show the model is **robust to foreseeable drift** and **monitored for unexpected drift.**

Below is a practical GMP-ready approach:

———

Step 1 — Establish Baseline Data Distributions

During validation, capture statistics such as:

- mean, median, and variance of inputs
- cluster structures
- feature correlations
- sensor baseline noise levels
- instrument signature patterns
- seasonal variations if relevant

This becomes your **reference distribution**.

Practical example:

For HPLC:

- retention time distribution
- peak width distribution
- baseline noise distribution
- detector signal shapes

Step 2 — Perform Stress Testing with Perturbations

Simulate realistic data drift by introducing:

- ± noise injection (2 percent, 5 percent, 10 percent)
- slightly shifted retention times
- varied instrument aging profiles
- brightness or contrast variations in EM images

Expected outcome:

Model remains stable. If small perturbations break the model → it is fragile and not credible.

Step 3 — Holdout "Future Likely Scenarios"

If you know a process change is coming (new media, new instrument), create a test set that represents the future state.

Example:
Introduce HPLC data from a different column model as part of validation.

Step 4 — Define Drift Thresholds

Define acceptable drift levels before validation is approved.

Example thresholds:

- population stability index (PSI) > 0.1 → investigate
- KS statistic shift > 0.2 → generate AI Monitoring Event
- principal component shift > 10 percent → evaluate
- feature mean shift > 2 standard deviations → flag

Step 5 — Build a SOP for Drift Monitoring

Validation must document:

- Who reviews drift
- How often (usually monthly)
- What metrics trigger action
- What actions occur (retrain, restrict use, escalate)

———

How to Monitor Drift in Production (GMP-Ready)

Below are realistic methods that auditors recognize.

1. **Distribution Monitoring (Data Drift Detection)**
   You track whether today's inputs statistically differ from baseline.
   Metrics:
   - PSI (Population Stability Index)
   - KL divergence
   - KS statistic
   - Chi-square tests
   - Feature mean/variance shift
   - Clustering changes

   Example alert:
   PSI = 0.13 for "detector baseline noise" → drift event created.

2. **Prediction Drift Monitoring**
   Even if inputs look normal, outputs may drift.
   Metrics:
   - distribution shift in predicted classes
   - spike in low-confidence predictions
   - higher analyst override rates
   - sudden shift in most common predictions

   Example alert:
   AI suggests "pump failure" 60 percent of the time this month vs. 22 percent historically.
   → Investigate process or model.

3. **Performance Drift Monitoring (Most Important)**

   You check whether the model is still **accurate.**

   - accuracy per failure mode
   - precision/recall
   - false negative rate
   - calibration error
   - agreements with SME review
   - comparisons to historical baselines

   Example alert:
   Model accuracy drops from 92 percent to 86 percent → retraining may be needed.

4. **Metadata Drift Monitoring** - Changes in:
   - instrument firmware
   - SOPs
   - operator behavior
   - batch sizes
   - raw material lots

   may silently cause drift.

5. **Edge Case Drift Monitoring**
   Track how the model handles:
   - rare events
   - anomalies
   - unusual chromatographic behavior
   - emerging microbial colony types

   These reveal concept drift sooner.

———

GMP Case Study: Drift in a Bioreactor Sensor Anomaly Detection Model

Scenario
A site uses an AI model that detects anomalies in bioreactor sensors (temperature, pH, DO, agitation).
Used for **decision support** only. **Validation accuracy: 94 percent.**

1. **What Drifts Happened**

   Data Drift - DO sensor readings started showing a wider variance because a raw material supplier changed.
   Concept Drift - As the bioreactor aged, agitator vibration noise created new patterns of spikes not seen in training data.

2. **How It Was Detected**
   Monthly monitoring log showed:
   - PSI for DO input = 0.18 (unexpected shift)
   - Increased number of low-confidence predictions
   - SME override rate increased from 6 percent to 15 percent

- Drift in PCA visualizations of raw signals
- Model repeatedly mislabeled "agitation jitter" as "oxygen depletion"

Drift confirmed.

3. **Actions Taken**
   - Deep dive root cause analysis
   - Retraining with:
   - new supplier lots
   - aged-equipment data
   - synthetic jitter augmentation
   - Configuration updated:
   - confidence threshold increased
   - narrow COU until retraining complete
   - Deployment of new model version controlled under QMS
   - SOP updated to reflect new drift threshold
4. **Outcome**

   Performance restored:
   - Accuracy increased back to 92 percent
   - SME override rate dropped to 7 percent
   - Drift indicators stable for 3 months

   Regulators accept this as proper drift management.

APPENDIX G: HUMAN OVERSIGHT & EXPLAINABILITY - REG EXPECTATIONS

When we validate AI in GMP, two concepts come up in every major regulatory document: **human oversight** (often called Human in the Loop, or HITL) and **explainability.** These two controls show inspectors that we understand the behavior of the AI, we can challenge it, and we can prevent incorrect outputs from influencing quality or safety.

These concepts are not optional. They are stated or implied across FDA, EMA, PIC/S, ISPE, and the EU AI Act.

FDA (2025, AI Draft Guidance)

FDA expects that any AI system supporting regulatory or quality decisions must maintain **appropriate human oversight** and allow users to understand "model behavior, limitations, and interpretability appropriate to the intended decision" (2025, FDA Draft Guidance).

EMA (2024, Reflection Paper)

EMA calls for a **human centric approach,** stating that users must be able to "review, verify, and override AI assisted outputs," and that systems used in higher-impact processes should provide **explainability proportional to risk** (2024, EMA Reflection Paper).

EMA (2025, Multidisciplinary Guideline on AI)

The updated guidance requires that AI systems "support human judgement" and provide transparency about how outputs were produced when those outputs influence quality, benefit-risk evaluations, or PV activities (2025, EMA Multidisciplinary AI Guideline).

PIC/S Annex 22 (2025 Finalization Update)

Annex 22 is the most explicit GMP reference. It requires:

- **human oversight mechanisms for high-impact uses, and**
- **explainable outputs that inspectors and personnel can understand** (2025, PIC/S Annex 22).

ISPE GAMP 5 Second Edition (2022/2023 AI Extensions)

GAMP defines HITL as a primary mitigation for nondeterministic or adaptive systems and recommends using **interpretable models or documented explanations** when AI impacts quality decisions (2023, ISPE GAMP 5 Updates).

FDA Good Machine Learning Practice (GMLP)

Although oriented toward medical devices, FDA lists **human factors and interpretability** as core principles needed to support safe use of AI systems (2021, FDA GMLP Principles). GMP inspectors often reference this.

ICH Q9 (Quality Risk Management)

While not AI-specific, Q9 establishes that human review is a formal **risk control**, and that improved detectability and transparency reduce process risk (ICH Q9). These concepts directly map to HITL and explainability.

EU Artificial Intelligence Act (2024–2025)

The AI Act requires "effective human oversight" and that users can **interpret and understand** the system's output for high-risk AI systems. While not GMP-specific, EMA aligns to it where appropriate.

Summary

Across all major regulators the direction is consistent:

- AI must not remove **human responsibility i**n GMP.
- AI decisions must be **traceable, understandable, and challengeable.**
- HITL and explainability are mandatory for any AI supporting quality, PV, or regulatory decisions.

These expectations shape how we scope validation, define requirements, and build risk controls in GMP environments.

2. Inspector-Facing Justification for HITL (for Your Validation Plan)

Below is a justification you can drop directly into the **Validation Strategy** or **Risk Management Approach** section of your AI Validation Plan. It is written as if you are explaining directly to an FDA or EMA inspector.

Inspector-Facing Rationale for Human in the Loop (HITL)

*The AI system in scope is used as a **decision support tool.** It provides suggestions or classifications, but it does not perform any automated actions or make final quality decisions. To ensure compliance with GMP expectations for nondeterministic technologies, we have implemented Human in the Loop (HITL) controls as a primary risk mitigation measure.*

HITL ensures that:

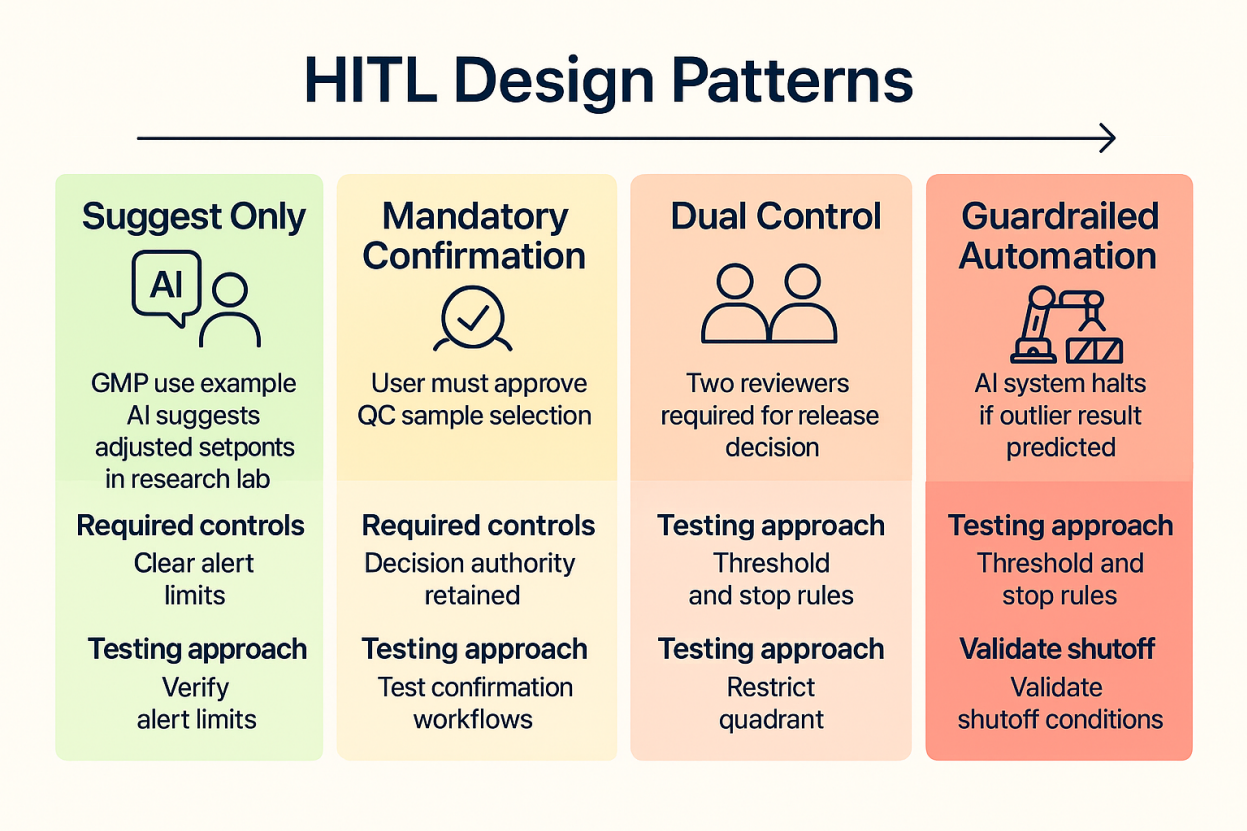
1. ***Qualified personnel review and confirm all AI assisted outputs** before they can influence any GMP decision or batch record. This is consistent with FDA's requirement for "appropriate human oversight" in AI systems supporting regulatory decision making (2025, FDA Draft Guidance).*
2. *The system maintains **human judgement and override capability**, aligning to the EMA Reflection Paper's requirement for a human centric approach and EMA AI Guideline expectations for high-impact decisions (2024, EMA Reflection Paper; 2025, EMA AI Guideline).*
3. *AI predictions remain **challengeable, traceable, and explainable,** consistent with PIC/S Annex 22 expectations that AI outputs must be explainable to users and inspectors (2025, PIC/S Annex 22).*

4. *HITL functions as a formal **risk control under ICH Q9**, increasing detectability and reducing the probability of an incorrect AI output impacting patient or product safety (ICH Q9).*
5. *The control design matches ISPE's GAMP 5 Second Edition guidance, which identifies HITL as a recommended mitigation for AI model uncertainty and non deterministic behavior (2023, ISPE GAMP Updates).*

In practical terms, this system requires a human reviewer to:

- *confirm or override all AI suggested outputs*
- *document their final decision*
- *provide justification where their decision differs from the AI*
- *ensure that any AI output is consistent with data, process knowledge, and GMP requirements*

Audit trail records include both the AI suggestion and the reviewer's final determination.
This approach ensures the AI system supports efficiency and consistency without replacing human expertise or GMP accountability. The system remains fully compliant with global regulatory expectations for human oversight in AI enabled processes.



APPENDIX H: REGULATORY LANDSCAPE (FDA, EMA, ISPE, PIC/S)

Before crafting any validation plan for an AI system, I always check the latest regulatory guidance. Why? Because agencies like the FDA and EMA have been actively thinking about AI and issuing guidelines to ensure these technologies are used safely in regulated environments. The good news is that the regulators' expectations **align with the core principles we already know** – things like risk management, documented evidence, and fitness for intended use – but they add AI-specific considerations (like needing robust data and monitoring model performance). In this section, I'll summarize the key points from major regulators and industry bodies:

FDA (U.S. Food & Drug Administration)

The FDA has been among the leaders in providing guidance on AI, particularly for medical devices and more recently for drug development and manufacturing. One cornerstone document is the **FDA's January 2025 draft guidance** titled *"Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products."* This draft guidance outlines a **risk-based framework for assessing the credibility of AI models** in any context where they support decisions on drug safety, efficacy, or quality. In plain language, FDA wants to ensure that if we use an AI model to support any regulatory decision (like deciding if a lot is within quality specs or if a certain process deviation is critical), we have evidence that the model is *credible* and reliable for that specific use.

The FDA's framework is a ***7-step process*** (2025, FDA Draft Guidance) that mirrors a lot of what we do in validation and quality risk management:

1. **Define the Question and Context of Use:** Clearly state what decision the AI will support and in what context it will be used (e.g., "This model predicts batch yield based on initial parameters to help release decisions"). This is akin to defining user requirements and intended use – it scopes what the AI will and *will not* do (appendix c).

2. **Assess Model Risk:** Determine how much risk is involved if the model's output is wrong. FDA breaks this down into factors like the model's influence on the decision (is it just advisory or automatic?) and the potential harm of a wrong prediction. If an AI only gives a recommendation that a human can override (like a decision support tool), that's lower risk than an AI making autonomous decisions. This risk assessment will guide how much validation rigor is needed.

3. **Plan to Establish Credibility:** Based on the risk, plan what evidence is needed to show the model is credible (accurate, robust, etc.). FDA calls this a "Credibility Assessment Plan" – basically our validation plan for the AI model. It might include what metrics we'll evaluate, how we'll test the model, acceptance criteria, and so on.

4. **Execute the Plan (Gather Evidence):** Carry out the testing and collect data on model performance according to the plan. For example, test the model on historical batches to see prediction accuracy, challenge it with edge cases, etc.

5. **Document Results and Deviations:** Document the outcomes, and if something didn't go as planned (maybe the model failed a certain scenario and you had to tweak it), note that and justify it. Sound familiar? It's basically the report of OQ/PQ for the model.

6. **Determine Adequacy for Intended Use:** Now decide if the model is "good enough" for the intended context. If the model met all acceptance criteria and risks are mitigated, you conclude it's adequate. If not, you might decide the model isn't ready or needs retraining/adjustment.

    E. Adequacy or "good enough" is defined as ***the model performs reliably, consistently, and safely within its defined operating domain for the intended use, with all known risks reduced to an acceptable level***. (See Appendix A for details)

    F. This is decided based on

        (1) The severity of harm if the model is wrong
        (2) The role of human oversight
        (3) The criticality of the decision
        (4) The strength of the validation evidence

7. **Lifecycle Management: A**lthough not an official "step" in the numbered list, FDA emphasizes you must maintain the model's credibility over time. That means monitoring performance in production, re-assessing risk if the context changes, and updating the model in a controlled way when needed (2025, FDA Draft Guidance).

One thing I appreciate in the FDA's guidance is it references the importance of **Good Machine Learning Practice (GMLP),** essentially applying quality system principles (design control, verification/validation, change control) to machine learning development. FDA had earlier published guiding principles for AI/ML in medical devices in 2021, and those ideas carry into current thinking: things like using good data, monitoring for bias, and ensuring transparency. For GMP professionals, the takeaway is that ***F*DA expects AI systems to be validated and maintained with a risk-proportionate approach, just like any other system, but also to account for data/model-specific risks.** If you validate an AI-based system for manufacturing or lab use, be prepared to explain to an inspector how you determined it was reliable (credible) for its specific purpose and how you're controlling it over time.

EMA (European Medicines Agency) and EU Guidelines

Over in Europe, there has been a flurry of activity to guide AI use in the pharmaceutical context. Two key documents stand out: the **2024 EMA Reflection Paper on AI** and the upcoming **2025 EU GMP Annex 22 on Artificial Intelligence.**

EMA's 2024 Reflection Paper

- This is a high-level document where EMA, along with the EU's Big Data Steering Group, reflects on how AI/ML can be used across the medicinal product lifecycle (from drug discovery through manufacturing to pharmacovigilance). It's not a binding guideline, but it outlines principles and concerns. For example, it stresses that *AI applications must comply with existing laws (GMP, GDPR, etc.) and uphold ethical and human-centric principles*. It acknowledges the great promise of AI to improve processes, but also warns that **"new risks are introduced that need to be mitigated to ensure the safety of patients and integrity of clinical study results"** (2024, EMA Reflection Paper). Key themes from the reflection paper include:
    - **Human-Centric Approach:** Always keep a human in the loop for critical decisions, and design AI with the end-user and patient in mind. In practice, for us, that might mean we don't blindly automate everything – we use AI to assist, not replace, human oversight where appropriate.
    - **Bias and Transparency:** EMA is clearly concerned about algorithmic bias and "black box" algorithms. They encourage seeking early advice if an AI could impact a medicine's benefit-risk so regulators can weigh in on your approach. As validation folks, we should document how we addressed bias (e.g., by using diverse training data, testing for biases) and how we'll ensure the AI's outputs are understandable or at least actionable.
    - **Data Quality:** Echoing what we already know from data integrity guidance, if an AI is trained on or processes data for regulatory use, that data had better be accurate and trustworthy. The reflection paper implies something straightforward: *garbage in, garbage out* – so data governance is as important as ever, if not more.
- The reflection paper paved the way for more concrete guidelines, and that's where Annex 22 comes in.

EU GMP Annex 22 (Draft, 2025)

- This is important because it's a dedicated annex in the GMP guidelines for AI and ML. It was developed jointly by EMA inspectors and PIC/S to ensure global alignment. As of mid-2025 it's in draft (consultation) form, but it signals very clearly how Europe expects AI to be handled in GMP environments. Here are the highlights (2025, EMA Draft Annex 22):
    - **Scope and Applicability:** Annex 22 applies to AI/ML models used in the manufacturing of medicines and active substances, **but only to certain types of models.** Notably, it says it covers *"static, deterministic AI/ML models"* and explicitly states that continuously self-learning (dynamic) models or models with probabilistic outputs **"should not be used in critical GMP applications."**

This means if you have an AI that updates its own algorithms on the fly or produces non-repeatable outcomes, regulators don't want that touching product quality or patient safety decisions without human oversight. Most likely, any AI you deploy in GMP should produce the same output given the same input, and any learning/update must go through a controlled retraining + validation cycle.

- **Intended Use & Data Characterization:** The annex puts strong emphasis on defining the *intended use* of the AI model in depth. You need to describe what task the model is performing, the process it's part of, and importantly, characterize the **data domain** – all the types of input data the model will handle, including variability and edge cases. For example, if you have an AI inspecting images of injectable vials, you must detail what kinds of vials, defects, lighting conditions, etc., are expected. Essentially, regulators want assurance that you understand the model's operating space and have considered the "known unknowns."

- **Performance Metrics and Acceptance Criteria:** Annex 22 requires companies to define **quantitative metrics** for model performance and set **acceptance criteria** *before* the model is tested/validated. This is analogous to setting acceptance criteria in a validation protocol – but here it might be things like "the model must identify at least 95% of defective units (sensitivity) and have no more than 1% false alarms (false positive rate) when tested on a representative dataset." Moreover, **the model's performance target should be at least as good as the manual or previous process it's replacing** (common-sense, but worth stating – you shouldn't deploy an AI that performs worse than the status quo) (2025, EMA Draft Annex 22).

- **Quality of Training and Test Data:** The annex calls for rigorous management of data used to train and test the AI:

  - All **test data** must be independent of training data (to avoid overfitting bias) and should be representative of the real use scenario. They expect you to document how you selected test data and why it's sufficient (covering common and rare cases).

  - Test data must be **accurately labeled** (e.g., if humans labeled images as "defective" or "good," there should be a verified process for that, possibly double-check by experts or using validated methods) – essentially ensuring ground truth is reliable.

  - Data preparation steps should be documented (if you normalized or augmented data, etc., to ensure you didn't inadvertently contaminate test data or skew results).

- **Validation and Testing Process:** Similar to our traditional approach but tailored:
  - The model needs to be tested (qualified) on the predefined acceptance criteria *using the independent test set*. This is your AI OQ/PQ equivalent. If it fails criteria, you investigate, adjust, or retrain with new data under change control and then re-test.
  - Annex 22 also touches on **explainability and confidence:** If the model provides confidence scores or probabilistic outputs, you should define how those will be used (for instance, will you only act on predictions above a certain confidence?). While they said not to use inherently probabilistic models in critical apps, even a deterministic model might give a confidence metric for each prediction – so clarify how that factors into decision-making.
  - All of this – model design, training, testing, and results – should be well documented and reviewed by QA. Even if you use a vendor-supplied AI solution, **you as the regulated user must have access to sufficient documentation** to justify the validation. I interpret this as: get documentation from your supplier about how the model was developed and tested, or perform your own testing, or both. You cannot just accept a "black box" without evidence.
- **Continuous Oversight & Change Control:** One of the most important parts of Annex 22 is that it **"foresees a continuous oversight of AI systems, including change control, model performance monitoring and procedures for human review when necessary."** In practice, after you put the model into operation:
  - **Change control:** The model (and its supporting software) should be under configuration management. If the model is retrained or updated, that's a change that should go through your change control system with appropriate impact assessment and re-validation.
  - **Performance Monitoring:** You need to monitor the model's performance at regular intervals or in real-time to ensure it's still meeting criteria. For example, track if error rates are creeping up or if incoming data starts to look very different from the training set (data drift).
  - **Data Drift Monitoring:** If the input data distribution shifts (say you start manufacturing a new product variant that the model wasn't trained on), the annex expects you to detect that. You might set statistical triggers – e.g., if the model starts getting a lot of "out of scope" inputs or if outcomes change significantly, that signals it's time to retrain or adjust.

- **Human-in-the-Loop Review:** If your AI is advisory (decision support), ensure that humans are actually reviewing outputs and that procedures exist for them to override or escalate issues. Annex 22 indicates that when AI reduces a human's effort in a task, you need procedures to ensure the human still remains vigilant to catch errors. For example, if an AI filters 90% of irrelevant deviation reports and only shows a human the likely critical 10%, you must train that human to trust but verify and have a process if they disagree with the AI or suspect it's missing something.

In summary, **EMA's Annex 22 (with PIC/S)** is setting a clear expectation: *treat AI with at least the same rigor as any GMP-critical system, plus manage data and model-specific challenges*. For us validators, it means more documents to produce (data specs, model test plans), more collaboration (with data scientists or vendors), and an ongoing commitment to monitor performance post-deployment. It might sound like a lot, but it's rooted in basic GMP principles: know what your system is supposed to do (intended use), make sure it actually does it (validation), and keep it under control (change management and monitoring).

PIC/S (Pharmaceutical Inspection Co-operation Scheme)

PIC/S works closely with EMA (and includes many global regulators, including FDA) to harmonize GMP guidance. In fact, as noted above, Annex 22 is a **joint EMA/PIC/S** initiative. So, everything described for Annex 22 applies to PIC/S member countries as well. PIC/S did not issue a separate different guideline; they are co-authoring and will adopt Annex 22 likely word-for-word to maintain alignment.

What's worth noting is PIC/S has been highlighting AI in its discussions. The *PIC/S Committee* recognized that AI is a major change in manufacturing tech, hence updating both Annex 11 (Computerised Systems) and creating Annex 22 for AI simultaneously. For you and me, if you're in a PIC/S member country (which includes many beyond Europe, like Australia, Canada, Japan, etc.), you can expect your inspectors to follow those same Annex 22 principles. A PIC/S news release summarized it nicely: *the new Annex 22 sets requirements for AI model selection, training, validation, focusing on defining intended use, establishing performance metrics, ensuring training data quality, and managing test data, with continuous oversight including change control and human review*. It's essentially the global inspector playbook for AI in GMP.

One more PIC/S angle: PIC/S published a guidance called PI 011 for Good Practices for Computerised Systems in regulated "GxP" environments and has discussed AI in forums, but nothing separate from Annex 22 as of 2025. So, we treat EMA and PIC/S as one voice on this topic for now.

ISPE GAMP 5 Second Edition (Industry Guidance)

Regulators aren't the only source of guidance – industry groups like ISPE have also updated their best practice frameworks to include AI. If you've worked in CSV, you're probably familiar with ISPE's **GAMP 5** (Good Automated Manufacturing Practice) guide, which is essentially a how-to manual for implementing risk-based validation. In **GAMP 5 Second Edition (2022),** ISPE added new content to address modern technologies including cloud computing, Agile development, and yes, AI/ML.

The GAMP 5 Second Ed. introduced an Appendix D11: *"Artificial Intelligence and Machine Learning"*, which **provides a risk-based life cycle framework for AI/ML, aligned with GAMP principles.** This appendix basically extends the familiar V-model and life cycle approach to cover the activities specific to machine learning systems. A few key points from GAMP worth noting:

- It emphasizes the idea that an AI/ML component can be seen as a *sub-system* within a larger system. For example, you might have a larger computerized system (like an environmental monitoring system) that includes an ML model as one component. GAMP suggests validating the overall system while paying special attention to that ML sub-system's life cycle.

- GAMP's AI appendix outlines life cycle phases similar to concept, project, operation – but includes data-specific and model-specific activities in each. For instance, in the **Concept** phase, you'd define the problem and evaluate feasibility of using AI, including consideration of data availability and algorithm selection. In the **Project** (design/development) phase, you perform model development iteratively: selecting algorithms, preparing data, training the model, and testing it, all under a plan that ties into your system requirements. In **Operation,** it stresses monitoring the model and data over time, similar to Annex 22's continuous monitoring concept.

- GAMP also touches on **Data Integrity** for AI. There's a cross-reference to a GAMP Good Practice Guide on "Data Integrity by Design" which had an appendix on AI/ML as well. The essence is that data powering AI must be high quality and integrity must be assured throughout the data's life cycle. We need to consider not just the final system's compliance, but the pipeline of data that flows into model training and predictions.

- Another concept from GAMP 5 second edition is **critical thinking over documentation** (the shift toward "Computer Software Assurance"). This still applies with AI: we should focus validation efforts based on risk and complexity. For example, if you have an AI model making GxP-critical decisions, you focus a lot of critical testing there, whereas an AI doing a non-critical task (like optimizing a schedule) might be validated in a lighter way. The risk-based approach is echoed by regulators too (both FDA and EMA mention applying QRM to AI).

As a CSV professional, I found the GAMP guidance useful because it translates regulatory expectations into practical approaches – it bridges the gap by saying "here's *how* you might implement validation for AI." For instance, GAMP suggests having a **"Data Management Plan"** as part of your validation deliverables to describe how you handle training and test data, something I hadn't included in traditional validation packages before. It also reassures us that existing validation stages (like user requirements, functional specification, verification testing, etc.) are still applicable – we just need to add AI-specific content to them.

Takeaway: Both regulators (FDA, EMA/PIC/S) and industry guides (ISPE GAMP) converge on the idea that **AI systems should be validated with a risk-based, life cycle approach, ensuring data quality, clear intended use, measurable performance, and ongoing control.** In the next sections, we will put these principles into practice. We'll walk through the AI system life cycle (so you know what activities to plan for) and then delve into risk management and concrete validation steps for typical use cases like decision support and lab automation.

(Side note: Other references like the FDA's **Predetermined Change Control Plan (PCCP)** *guidance for AI-enabled medical devices, or the upcoming EU AI Act, are also part of the broader landscape. However, for a GMP CSV focus, the documents we covered are the most directly relevant. The AI Act is EU-wide legislation classifying AI systems by risk – if you're using AI in GMP, it likely falls under high-risk, requiring robust risk management – which, as you see, is exactly what Annex 22/GAMP are prescribing.)*

About the Author

Michael Weaver

AI-Enabled Data Integrity, Regulatory Intelligence, and Risk Leadership for Biopharma

Mike is a senior quality and data governance leader with experience strengthening inspection readiness, digital compliance, and enterprise risk management across global biotech operations. His work includes Part 11 modernization, ISO 13485/14971 alignment, cross-functional risk governance, and AI-driven compliance automation. He has trained more than 1,200 staff worldwide and has led programs that contributed to first-time FDA approvals.

He focuses on practical, risk-based adoption of AI in GMP environments, helping teams validate AI systems, modernize documentation workflows, and build resilient digital quality frameworks.

Email: michael.weaver@qualprep.com

LinkedIn: linkedin.com/in/mikeweaver2

Website: https://www.qualprep.com