ChatGPT

Validating AI Systems in Pharma

1. Validation Fundamentals

In pharmaceutical manufacturing, **validation** ensures that processes, equipment, and systems consistently produce quality products as intended [1]. Validation covers various domains: *process validation* (ensuring a manufacturing process yields reproducible results), *cleaning validation* (demonstrating cleaning procedures remove residues to acceptable levels), and *analytical method validation* (proving lab test methods are accurate, precise and suitable) [1]. **Computer System Validation (CSV)** is the application of validation principles to computerized systems, ensuring software (e.g. control systems, databases) performs reliably and securely.

Qualification vs. Validation: Equipment qualification (IQ/OQ/PQ) is part of the overall validation framework. Qualification (IQ/OQ/PQ) provides "documented evidence that equipment or systems are fit for their intended use" [2]. In other words, qualification verifies the equipment itself (installation, operation, performance). Validation is broader: it encompasses the entire process, including these qualified equipments plus materials, methods and controls, to ensure consistent product quality [1]. Thus, qualification is an input to process validation, laying the foundation for a validated process.

2. Why AI Creates New Validation Challenges

AI-based tools introduce unique complexities for validation:

- **Non-Deterministic Behavior:** Traditional systems respond predictably to given inputs. AI (especially machine learning models) may produce different outputs for the same input if not fully deterministic (e.g. due to random elements or parallel processing). This unpredictability challenges the reproducibility expected in GMP.
- **Continuous Learning and Model Drift:** Some AI models learn and update from new data over time. While beneficial in production, continuous learning means the model's performance can change post-qualification. Detecting and controlling *model drift* is hard, as even the same inputs may yield evolving results without obvious changes in software version.
- **Explainability Limits:** Many AI algorithms (e.g. neural networks) operate as "black boxes." The rationale for a given output is not transparent. Regulators demand that validated systems be traceable and explainable. If an AI system makes a recommendation, QA must understand and justify it, which is difficult if the model cannot explain its reasoning [3].
- **Dependency on Data Quality:** AI performance depends heavily on the quality and representativeness of its training data. Biased, incomplete or poor-quality data can lead to unreliable AI outputs. During validation, one must verify not just the code, but also the data pipeline and pre-processing.
- **Reproducibility Issues:** Even if an AI model is fixed, small changes (like different hardware or library versions) can yield slight output variations. This undermines the exact reproducibility principle of validation.

- **Configurable vs Adaptive Systems:** Many AI tools have tunable parameters or self-adaptive features. Each configuration change can alter behavior. Validation must account for possible settings, leading to a combinatorial explosion of test cases if not well controlled.

In summary, AI's "continuous learning and opacity" run counter to GMP's emphasis on fixed, reproducible systems [3]. These characteristics make traditional validation approaches insufficient without adaptation.

3. Regulatory and Industry Expectations

Regulators currently apply GxP principles to AI systems with an emphasis on risk and documentation:

- **Intended Use:** As with any tool, an AI system's intended use must be clearly defined. This determines the validation scope. For example, an AI used for simple data filtering is inherently lower risk than one making pass/fail decisions on a batch.
- **Risk-Based Validation:** FDA and EMA guidance encourage a risk-based approach to AI [4]. Low-impact AI functions (e.g. scheduling optimization) require minimal oversight, while high-impact AI (e.g. real-time process control affecting product quality) demand comprehensive validation [4] [5]. Table 4 of a recent review illustrates this classification by impact and complexity [4] [5]. For example, predictive maintenance alerts (low to medium risk) need basic validation and monitoring, but AI controlling sterile operations (critical risk) would need full regulatory pre-approval.
- **Documentation:** Every aspect of the AI system must be documented per GMP. This includes development details (algorithm architecture, training data, hyperparameters), performance metrics, and rationale for design choices [6]. FDA/EMA expect version control and audit trails for AI models just as for software or instruments. Emerging tools can automate generating compliance documentation from model development [6].
- **Change Management:** Any update to an AI model (retraining, new algorithms, data changes) is akin to a "change control" event. A controlled change procedure is needed, including impact assessment and revalidation triggers. The FDA's AI discussion paper (for medical devices) already recommends a total product lifecycle approach, which would apply similarly to manufacturing AI.
- **Performance Monitoring:** Post-deployment, AI tools require ongoing monitoring. Just as environmental monitoring is done for equipment, model performance should be tracked (e.g. accuracy, error rates) to detect drift. Any degradation necessitates investigation and possibly revalidation.
- **User Oversight:** Regulators emphasize that AI should augment, not replace, human expertise [7]. Qualified personnel must oversee AI outputs, with the ability to intervene. The concept of "human-in-the-loop" (explicit human decision) or "human-on-the-loop" (monitoring) is being formalized in guidance [7].

In essence, existing GxP principles apply: AI systems must be validated to prove fitness for use under their intended context. This is typically done with a risk-based plan, strong documentation, change control, and evidence of effective human oversight [6] [7].

4. Validation Lifecycle for AI-Enabled Tools

Validating AI tools generally follows a tailored lifecycle:

1. **Define Requirements and Use Case:** Clearly document the AI tool's intended use, inputs, and expected outputs. Specify accuracy/performance requirements and acceptable error margins for its context.
2. **Data Preparation and Training:** Document sources and quality of training data. Perform data qualification (e.g. ensuring calibration of sensors, lab instruments supplying the data). Capture the entire training process.
3. **Model Development and Selection:** Keep records of algorithm choices, architectures, and training runs. Use explainability tools where possible to understand model behavior.
4. **Challenge Testing:** Test the AI on known validation datasets (distinct from training data) to assess performance. Include boundary testing (extreme conditions) and adversarial cases to ensure robustness.
5. **System Integration Testing:** Verify the AI system as a whole: inputs flow correctly, outputs are captured, and user interfaces work. Ensure audit trails of data flow.
6. **User Qualification:** Train users on system operation and limitations. Document user qualifications similarly to software or instrument training.
7. **Verification of Outputs:** Compare AI outputs with independent assessments or known standards. For example, if AI is used to classify samples, verify with lab results. Any discrepancies must be investigated.
8. **Documentation Package:** Assemble a formal report of the above: validation protocols, test results, deviation reports, and final qualification.
9. **Deployment and Monitoring:** After approval, monitor the AI system in production. Define control charts or alerts for key performance indicators. For instance, track prediction accuracy over time.
10. **Revalidation Triggers:** Establish criteria to revalidate the model: e.g. major software update, significant change in process, or observed model drift beyond threshold.

This lifecycle resembles traditional CSV but includes AI-specific steps (like retraining tracking and model performance monitoring). The key is ensuring reproducibility and traceability at every stage, even though the model itself may adapt.

5. Use-Case Categories (Risk Levels)

AI applications can be categorized by risk level:

- **Low-Risk Uses:** Functions like drafting document summaries, providing suggestions for SOP language, or analyzing historical data for descriptive reports. These are advisory roles; a human must review all outputs. Risk is minimal since final decisions rest with the user.
- **Medium-Risk Uses:** Examples include quality analytics dashboards, trend analysis of deviations, or predictive maintenance alerts. These support decision-making. AI can recommend actions (e.g. schedule maintenance) but a person verifies the recommendation. Basic validation and monitoring are needed.
- **High-Risk Uses:** Decision support tools that influence critical outcomes, such as real-time adjustment of process parameters or batch release recommendations. AI here directly affects

product quality. This requires rigorous validation (perhaps formal statistical evidence) and thorough documentation.
- **Critical Uses:** Fully automated or "closed-loop" control systems (e.g. AI controlling reactor conditions in a sterile fill line). These are safety-critical. Regulatory agencies would expect such systems to undergo full validation and likely separate regulatory evaluation.

This risk gradation is echoed in regulatory literature [4] [5]. Generally, the higher the risk to product quality or patient safety, the deeper the validation and human oversight required.

6. Risks

Several specific risks accompany AI in this context:

- **Hallucinations and Inaccuracy:** Generative models (LLMs) can produce fluent but factually incorrect outputs. Using such outputs for quality reports or labels without verification can be dangerous. Users must critically review all AI-generated content.
- **Instability:** AI algorithms might give different results on subsequent runs due to randomness or model updates, undermining consistency.
- **Hidden Model Updates:** Cloud-based AI services may silently update underlying models. If a supplier changes the model version, it may alter behavior unexpectedly. Without transparency, this is a serious validation gap.
- **Vendor Transparency:** Some AI platforms are proprietary "black box" solutions. If a vendor won't disclose key details (algorithm type, training data provenance), it is hard to validate and may violate the transparency regulators expect.
- **Insufficient Evidence:** Traditional validation relies on objective acceptance criteria and test records. AI outputs may be probabilistic scores. Establishing firm pass/fail criteria can be nontrivial. For example, how do you "approve" a chatbot's answer objectively?
- **Over-reliance:** There's risk that users may trust AI beyond its intended scope. If users assume an AI suggestion is authoritative without question (especially for black-box decisions), this could lead to compliance breaches.
- **Data Integrity:** AI systems often ingest large datasets. Ensuring data integrity (accuracy, completeness) is critical; faulty input data will produce faulty outputs. GxP regulations apply to this data pipeline as well.

Ultimately, these risks mean that any AI-enabled tool needs robust controls: audit logs, versioning, and human checks. It should be treated as a high-risk software in terms of validation rigor.

7. Top AI Platforms for Regulated Environments

Below are three AI platforms/tools noted for their enterprise governance and suitability for regulated use. Each offers features to support validation and control.

Platform	Best Use Case (Validation Focus)	Strengths	Weaknesses	Regulated Suitability	Governance Features
Microsoft Azure AI (OpenAI Service)	Research, document summarization, Q&A in controlled settings	*Enterprise-grade* with Azure compliance standards. Integrates OpenAI models with Azure Monitor, Key Vault, private endpoints. Offers GPT-4 with built-in safety layers.	LLM outputs can hallucinate; must restrict private data. Requires configuration to meet 21 CFR 11 (audit logs, access control). Dependency on cloud service updates.	High. Azure provides compliant infrastructure (FedRAMP, ISO, etc.) and features for logging, encryption, access control. Validation is more about controlling usage and data rather than model logic.	Role-based access, logging of model calls, content filtering tools. Azure OpenAI Service allows containerization and restricted networks.
IBM Watsonx	Data analytics, risk modeling, NLP-assisted reporting	*Explainable AI*: includes tools for model interpretability (AI Explainability 360). Enterprise data governance (IBM's established analytics portfolio).	Primarily big-data analytics, not specialized in generative chat. Onboarding complexity. Cloud or on-prem options.	High. IBM platforms have strong security/ compliance history. Watsonx adds enterprise features (audit trails, data lineage, explainability) which help in validation evidence.	AI FactSheets for model transparency, automated documentation, integration with IBM OpenScale for continuous monitoring.

Platform	Best Use Case (Validation Focus)	Strengths	Weaknesses	Regulated Suitability	Governance Features
DataRobot	Predictive modeling with automated compliance docs	*Automated MLOps*: generates documentation for models (data lineage, training logs) which aids validation. Supports model explainability (SHAP, etc.). Workflow tracking.	Focused on model development (not as much on LLM). Licensing and cost can be high. Cloud/Hybrid deployment.	Designed for regulated industries (banking, pharma). Provides compliance tooling and audit logs. Validating a model includes auto-generating the evidence pack.	Automatic documentation templates, audit trails for model changes, data preprocessing tracking. Ability to "lock" models once validated.

- **Microsoft Azure AI (OpenAI Service):** This lets companies use GPT-4/GPT-3.5 within Azure's controlled environment. It is relatively mature for enterprise use. For validation, Azure's compliance (e.g. SOC 2, HIPAA, ISO) helps ensure infrastructure controls. Role-based access and logging help meet 21 CFR 11. The main challenge is controlling the AI model: e.g. freezing to a specific model version and auditing usage. Use case: internal document drafting or Q&A where outputs are closely reviewed. Strengths include large model capabilities; weakness is lack of inherent explainability in LLMs.
- **IBM Watsonx:** IBM's AI platform emphasizes trust and explainability. Watsonx Assistants can do NLP tasks, and tools like AI Explainability 360 offer insights into model decisions. Data and model lineage is tracked. It supports hybrid (on-prem) deployment, which suits highly regulated sites. A use case is NLP analysis of quality records with traceable model decisions. Strengths are enterprise governance, but it requires skilled setup.
- **DataRobot:** A leading automated ML (AutoML) platform. It can train predictive models and automatically produce a "compliance documentation package" for each model. This package includes data summaries, model metrics, and even code snippets to aid validation. Useful for building predictive analytics (e.g. CAPA recurrence risk) with audit trails. Its main limitation is that it's focused on structured predictive models, not unstructured LLM tasks.

These platforms were chosen for their track records in enterprise and regulated contexts. All three support features like audit logging, version control, role-based access, and model explainability that align with CSV needs. However, none inherently solve the "validation of learning" problem – rather, they provide tools to document and monitor AI models as rigorously as possible.

8. Practical Framework for QA/CSV Assessment

When evaluating an AI tool for use, QA and CSV professionals should apply a structured framework:

1. **Define Context and Risk:** Determine the AI tool's intended use and impact level (see Section 5). Low-risk utilities need simpler controls; high-risk applications need full validation.
2. **Vendor Assessment:** Choose solutions with strong governance. Ensure the vendor provides transparency (model cards, documentation generation) and that you can control updates.
3. **Validation Plan:** Develop a validation plan analogous to CSV, with AI-specific elements: define performance requirements, testing datasets, acceptance criteria for outputs (e.g. accuracy thresholds), and methods for monitoring drift.
4. **Documentation:** Generate evidence bundles for the model (training data description, model version, hyperparameters, test results) – platforms like DataRobot can automate this [8].
5. **Change Control and Monitoring:** Integrate the AI tool into change control. Any model retraining or software update triggers impact assessment. Establish key performance indicators (KPIs) for ongoing monitoring (e.g. error rate vs baseline) and set limits for revalidation.
6. **Human Oversight:** Clearly define how subject-matter experts will review AI outputs. Maintain manuals and training on the tool's proper use and limitations.
7. **Periodic Review:** Schedule periodic re-evaluation of the AI system. For example, annually assess whether the model is still fit-for-purpose, whether data distribution has changed, and whether retraining or requalification is needed.

By applying the same rigor as traditional validation – with adaptations for AI's dynamic nature – organizations can responsibly leverage AI tools. The key is to treat AI models as configurable computerized systems: document every step, control changes tightly, and continuously verify performance. This allows AI to be a validated part of pharmaceutical quality systems, balancing innovation with compliance.

Sources: Regulatory guidance (FDA, EMA, ISPE GAMP5) and recent literature on AI/ML in GxP were reviewed. Key points on AI validation challenges and frameworks were drawn from a 2025 review of AI/ML in GMP environments [3] [4]. Information on platform governance (DataRobot, Azure, IBM) comes from vendor documentation and case studies [8].

---

[1] [2] Validation Qualification | Life Sciences industry | Cleanrooms
https://www.gxpcellators.com/validation-vs-qualification/

[3] [4] [5] [6] [7] Regulatory Perspectives for AI/ML Implementation in Pharmaceutical GMP Environments - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC12195787/

[8] Compliance: DataRobot docs
https://docs.datarobot.com/en/docs/workbench/compliance/index.html