Journal on
Artificial Intelligence

Tech Science Press

Doi:10.32604/jai.2026.073895

ARTICLE

Building Regulatory Confidence with Human-in-the-Loop AI in Paperless GMP Validation

Manaliben Amin*

Pharmaceutical Manufacturing, Stevens Institute of Technology, Hoboken, NJ 07030, USA

*Corresponding Author: Manaliben Amin. Email: manaliamin3@gmail.com

Received: 28 September 2025; Accepted: 04 December 2025; Published: 06 January 2026

ABSTRACT: Artificial intelligence (AI) is steadily making its way into pharmaceutical validation, where it promises faster documentation, smarter testing strategies, and better handling of deviations. These gains are attractive, but in a regulated environment speed is never enough. Regulators want assurance that every system is reliable, that decisions are explainable, and that human accountability remains central. This paper sets out a Human-in-the-Loop (HITL) AI approach for Computer System Validation (CSV) and Computer Software Assurance (CSA). It relies on explainable AI (XAI) tools but keeps structured human review in place, so automation can be used without creating gaps in trust or accountability. Within this framework, AI provides draft recommendations such as mapping user requirements, highlighting redundant tests, or classifying deviations while subject matter experts (SMEs) review, adjust, and approve the final outcomes. Features like rationale cards, confidence bands, and traceable recommendations give SMEs the information they need to understand and, when necessary, challenge AI outputs. The framework is also supported by governance measures. These include model risk tiering to match oversight with potential impact, periodic model challenges to detect drift, and "evidence packs" that bring together AI outputs, human decisions, and audit trails into an inspection-ready format. Taken together, these safeguards show inspectors that automation is being used responsibly, not recklessly. We validate the approach in a two-project pilot, reporting mean cycle-time reductions of 32% (95% CI: 25%–38%) and higher inter-rater agreement (κ from 0.71 → 0.85), achieved under a defined governance model that includes model-risk tiering, quarterly challenge testing for high-risk models, and inspection-ready evidence packs aligned to 21 CFR Part 11 and Annex 11. These results provide preliminary empirical support so that HITL AI can improve efficiency and reviewer consistency while preserving accountability and regulatory trust. Practical case examples demonstrate that HITL AI can shorten validation cycles by 25%–40% while also improving reviewer consistency and strengthening inspector confidence. Rather than replacing SMEs, the system frees them from repetitive work so they can focus on risk and quality the areas where their judgment adds the most value. By blending automation with accountability, HITL AI provides a path for digital transformation that regulators are more likely to accept, positioning it as both a productivity tool and a model for sustainable compliance in the years ahead.

KEYWORDS: Artificial intelligence (AI); Computer Software Assurance (CSA); Computer System Validation (CSV); Explainable AI (XAI); Good Manufacturing Practice (GMP) compliance; Human-in-the-Loop (HITL); paperless validation; pharmaceutical manufacturing; regulatory trust

1 Introduction

Pharmaceutical manufacturing is one of the most tightly controlled industries in the world. Every batch of medicine depends on reliable systems, and those systems must be thoroughly validated before they can be trusted. For a long time, this work was almost entirely paper-based. Teams filled binders with protocols, test

results, and signatures. The process worked, but it was slow, vulnerable to mistakes, and difficult to manage once systems became more complex. The move toward paperless validation platforms has eased many of these problems. By digitizing workflows, companies gained automation, searchable audit trails, and easier collaboration across departments. Recently, a new layer of capability has begun to appear in these platforms: artificial intelligence. Some tools can now generate draft test cases, highlight areas where testing effort could be reduced, or scan requirements and suggest possible test mappings.

These innovations promise clear efficiency gains. Still, regulators are cautious. They have raised concerns about "black box" systems that make decisions without explanation. Inspectors regularly point out that responsibility for quality and patient safety rests with people, not machines. Ethics researchers echo this point, arguing that AI should never replace human accountability but instead be designed to support it [1].

Human-in-the-Loop (HITL) AI provides one way to meet both goals. In this approach, AI generates suggestions, but subject matter experts (SMEs) always review and decide the outcome. Far from replacing expertise, the system speeds up routine work while leaving control in the hands of qualified staff. This model addresses the concerns of regulators and, at the same time, gives companies a practical path to benefit from automation.

Despite growing interest, there is limited empirical evidence quantifying the impact of explainable, Human-in-the-Loop AI on validation cycle-time, review quality, and consistency within GMP environments. Existing publications focus largely on conceptual discussions or regulatory summaries without providing reproducible, data-driven validation outcomes.

This paper aims to (i) operationalize an auditable HITL AI framework that integrates explainability artifacts, governance structures, and defined human decision rights, and (ii) evaluate its measurable impact through a two-project pilot with predefined KPIs and confidence intervals.

2 From Checklists to Risk-Based Thinking: Evolving Standards

Pharma companies deal with layers of international rules that cover how computerized systems are built, tested, and kept up to standard. These rules are not just bureaucracy. They exist to keep patients safe, make sure products meet quality expectations, and ensure the data used in decisions can be trusted.

One recent shift worth noting came in 2022, when the FDA released its draft guidance on Computer Software Assurance (CSA). The shift is important. Instead of asking teams to tick boxes and fill binders with repetitive tests, the guidance tells companies to concentrate on the checks that matter most those tied directly to product quality and patient safety [2].

In Europe, Annex 11 of EudraLex Volume 4 takes a slightly different approach but comes back to the same point. It demands accountability and traceability. Inspectors don't just want to see that validation happen; they also want proof of who signed off, when the approval was made, and why. At the end of the day, responsibility lies with people, not with the system. Globally, ICH Q9 (R1) builds on this by treating validation as a risk exercise. A single assessment at the start isn't enough. Risks have to be tracked and managed across the full lifecycle. That means applying proportional testing, writing down the reasoning when test scope is reduced, and continuing to monitor as systems and environments evolve. A summary of these major regulatory frameworks and their implications for Human-in-the-Loop (HITL) AI is provided in Table 1.

Finally, the ISPE GAMP 5 (Second Edition), updated in 2022, is often called the validation playbook. It supports digital tools, automation, and even AI but only if strong governance and documentation are in place. Regulators may welcome innovation, but not at the expense of trust or traceability.

Table 1: Key regulatory frameworks for computerized systems

Guideline /Framework	Region/Issuer	Key focus area	What it means for HITL AI in validation
FDA CSA Draft Guidance (2022)	United States	Pushes critical thinking and risk-based testing instead of paperwork-heavy approaches	Supports reducing unnecessary tests if the logic is clear and SMEs remain the final reviewers
EU Annex 11 (2011)	European Union (EMA)	Strong on accountability, traceability, and data integrity	Requires every AI suggestion to be linked to human approvals and full audit trails
ICH Q9(R1) (2023)	International Council for Harmonisation	Risk management across the system lifecycle	Calls for continuous monitoring and justification for reduced testing; AI can help, but SMEs must confirm
ISPE GAMP 5, 2nd Edition (2022)	Industry best practice (global)	Lifecycle-based validation with an eye on innovation	Endorses digital tools, including AI, if they are transparent, documented, and properly governed

Taken together, these frameworks converge on the same regulatory intent risk-based validation, proportional testing, and traceable human accountability. Within this context, Human-in-the-Loop AI can be positioned as a decision-support layer that augments but never replaces SME oversight. It preserves accountability through explicit decision rights, e-signatures, and complete audit trails, while mapping governance artifacts such as RACI matrices, periodic challenge tests, and inspection-ready evidence packs directly to these frameworks.

Regulatory agencies are also beginning to formalize expectations for AI governance. The FDA's *AI/ML Action Plan* (2023) and the EMA's *Reflection Paper on AI in the Medicinal Product Lifecycle* (2024) both emphasize transparency, model-risk management, and human oversight. Complementary industry analyses highlight validation challenges for AI-containing products in regulated manufacturing and outline GxP-aligned machine-learning governance models. Integrating these principles within the CSA and GAMP ecosystem situates the proposed HITL AI framework squarely within emerging global oversight trends.

Taken together, the message is consistent: digital transformation is possible, but it has to be done responsibly. Inspectors want to see clarity, accountability, and evidence that automation is under control. AI can speed things up, but it will never replace the oversight and judgment of SMEs.

3 Materials and Methods

3.1 Study Context

Two independent GMP digital-validation projects were evaluated to assess the impact of Human-in the-Loop (HITL) AI on validation efficiency and consistency. Both projects were executed on a commercial paperless-validation platform (Kneat Gx) augmented with a prototype HITL AI module. The first project

focused on user-requirement-to-test (URS→test) mapping, while the second centered on deviation classification and severity assessment. Each project represented an end-to-end validation lifecycle including authoring, review, and approval stages.

The combined dataset comprised approximately 650 user-requirement statements and 120 deviation records drawn from legacy anonymized validation reports. All data were de-identified prior to model ingestion, and access was restricted under role-based access control (RBAC) to preserve confidentiality and data integrity [3]. All AI models and workflows were operated under a controlled sandbox environment, ensuring no impact on live GMP records or production data.

3.2 Model Architecture and Explainability Components

Three complementary algorithmic modules were deployed:

Natural-Language Processing (NLP) mapping model: A BERT-based semantic-similarity engine identified potential test cases corresponding to each URS by computing contextual embeddings and cosine similarity scores [4]. Candidate pairs exceeding a 0.75 similarity threshold were shortlisted for SME review.

Deviation-classification model: A gradient-boosted decision-tree ensemble predicted deviation severity (minor, major, critical) using historical attributes such as affected system type, deviation description length, product impact, and recurrence frequency.

Explainability layer: Each model was wrapped with post-hoc explainability tools using SHAP (Shapley Additive Explanations) values to produce rationale cards. These cards listed the top influencing features contributing to each recommendation. Confidence bands (±σ on probability outputs) quantified uncertainty, while counterfactual examples showed how minimal feature changes could alter the outcome. This combination provided interpretable context for human reviewers and enabled downstream traceability during audits.

3.3 Human-in-the-Loop Governance Protocol

The HITL AI workflow ensured that automation functioned strictly as a decision-support tool, consistent with the oversight expectations outlined in the FDA AI/ML Action Plan (2023) [3] and the EMA Reflection Paper on AI (2024) [5].

AI generation: The model generated draft recommendations (e.g., URS–test mappings or deviation classifications).

SME review: Subject-matter experts (SMEs) evaluated each recommendation, selecting *Accept*, *Modify*, or *Reject* while adding justification comments.

QA verification: Quality Assurance (QA) independently reviewed all *high-risk* model outputs (those with confidence < 0.8 or Tier 3 risk category) to ensure appropriateness and compliance.

Approval and traceability: Every action was captured through e-signatures and automatically time-stamped in accordance with 21 CFR Part 11 and Annex 11 requirements. The resulting audit log contained the AI suggestion, SME decision, reviewer notes, and system-generated unique identifier.

Governance was supported by a RACI matrix delineating responsibilities (Validation Engineer–Responsible, SME–Accountable, QA–Approver, IT–Consulted). Challenge testing and periodic model-performance monitoring were conducted quarterly for high-risk models and semi-annually for moderate-risk ones to verify sustained reliability.

3.4 Key Performance Indicators (KPIs) and Statistical Treatment

Performance assessment combined quantitative KPIs with qualitative SME feedback to evaluate the effect of Human-in-the-Loop (HITL) AI on validation efficiency and reviewer consistency. Each KPI was defined at the **validation-cycle level**, where one cycle represents a complete authoring–review–approval sequence within a project plan.

- **Cycle-time** and **mapping-time** were normalized per user-requirement-set (URS set) with denominators of n = 12 baseline cycles and n = 12 HITL cycles.
- **Review/redo rate** was computed as the proportion of validation packages requiring partial rework after initial submission.
- **Inter-rater agreement(κ)** was calculated from 120 deviation-classification records benchmarked against QA consensus to measure reviewer consistency.

All continuous variables are reported as mean ± 95% confidence intervals obtained through **paired bootstrap resampling (n = 10,000)**, which preserves within-project correlation between baseline and HITL conditions. To control for testing multiple endpoints, a **Holm–Bonferroni family-wise error correction (α = 0.05)** was applied, and both absolute differences (Δ) and relative percentage changes are presented with corresponding adjusted p values.

Ablation analyses were designed to isolate the contribution of specific explainability and oversight components. Each ablation used random subsamples of 60 URS mappings and 30 deviation records per condition, drawn from the same dataset. Confidence-band thresholds were pre-defined at 0.75 (medium risk) and 0.85 (low risk) probability levels, as specified in the internal governance SOP HITL-GOV-002. In "no-rationale" ablations, the SHAP-based feature-attribution summaries were hidden from reviewers; in "no-confidence-band" tests, uncertainty intervals were suppressed while point estimates remained visible; and in "AI-only" exploratory runs, SME checkpoints were temporarily disabled inside the sandbox environment. These controlled experiments ensured that model performance could be compared both with and without human oversight, consistent with good-machine-learning-practice recommendations for regulated environments [6].

All model configurations, parameter files, and governance workflows were **version-controlled under Git repositories** within the validation sandbox. Each AI model maintained a **Model Master Record (MMR)** capturing training-data lineage, hyper-parameter settings, validation metrics, and quarterly challenge-test results. Independent validation teams replicated KPI calculations on a held-out 20% subset to confirm reproducibility. Confounding variables team composition, project complexity, and reviewer load were held constant across baseline and HITL conditions to ensure that observed improvements could be attributed to the Human-in-the-Loop framework rather than external factors.

3.5 Reproducibility and Controls

Model configurations, parameter settings, and governance workflows were version-controlled using Git repositories within the validation sandbox. Each AI model had a Model Master Record (MMR) documenting training-data lineage, performance metrics, and challenge-testing results. Independent validation teams replicated KPI measurements on a held-out subset (20%) to confirm reproducibility. Confounding variables such as team composition, project complexity, and review load were held constant across baseline and HITL conditions.

4 Results and Evaluation

4.1 Aggregate KPI Outcomes

The key results are summarized in Table 2, which compares baseline and HITL AI performance metrics with corresponding 95% confidence intervals and adjusted *p* values. Each KPI was defined at the validation-cycle level, where one cycle represents a complete authoring–review–approval sequence. Cycle time and mapping-time were normalized per URS set ($n = 12$ baseline, $n = 12$ HITL). Deviation-classification consistency (κ) and review/redo rate were calculated from 120 deviation records across both projects. All confidence intervals were generated through **paired bootstrap resampling (10 000 iterations)** to preserve within-project correlation, and **Holm–Bonferroni correction (family-wise α = 0.05)** was applied to control multiplicity. Effects are expressed as absolute differences (Δ) and relative percentage changes with 95% CIs.

Table 2: Key Performance Indicators (KPIs) comparing baseline and HITL AI validation performance

KPI	Denominator (*n*)	Baseline (Mean)	HITL AI (Mean)	Δ (Absolute)	Δ (%)	95% CI	*p* (Holm–Bonferroni)
Cycle-time (days)	12 cycles per arm	15.0	10.2	−4.8 days	−32	(−38, −25)	0.021
Mapping time (hours per URS set)	12 cycles per arm	10.0	6.7	−3.3 h	−33	(−40, −25)	0.018
Review/redo rate (%)	12 cycles per arm	21	13	−8 pp	−38	(−44, −28)	0.033
Inter-rater agreement (Cohen's κ)	120 deviation records	0.71	0.85	+0.14	—	(+0.08, +0.16)	0.027

Notes: One cycle = complete authoring–review–approval sequence. Cycle-time and mapping-time computed per URS set ($n = 12$ baseline; $n = 12$ HITL). κ and redo rate derived from 120 deviation records benchmarked against QA consensus. 95% CIs derived from **paired bootstrap** (10 000 iterations), percentile method; pairing preserves within-project correlation. *p* values adjusted for four primary endpoints using **Holm–Bonferroni** family-wise error correction (α = 0.05). "pp" = percentage points. Negative Δ indicates improvement (where lower is better).

Improvements in authoring and review efficiency are statistically significant at $p < 0.05$ (bootstrap based confidence intervals). The cycle-time reduction of roughly one-third and the 33% decrease in mapping-time demonstrate measurable efficiency gains without compromising traceability or audit readiness. The review/redo rate fell by about 38%, confirming fewer downstream corrections and more stable document quality. The increase in κ from 0.71 to 0.85 indicates stronger reviewer alignment, confirming that the explainability and structured-oversight mechanisms reduced subjective interpretation between SMEs. These results mirror time-saving effects reported in other AI-assisted quality-management applications and align with the efficiency targets promoted under the FDA CSA guidance.

Ablation analyses provided additional evidence that both explainability and human oversight are essential. When rationale cards were suppressed, SME acceptance of AI recommendations declined by ≈11 percentage points (95% CI 6–15); hiding confidence bands increased override frequency by ≈7 points (95% CI 3–11); and running the model in AI-only mode raised rework rates by ≈9 points while reducing κ to 0.74. These patterns confirm that automation alone introduces inconsistency, whereas Human-in-the-Loop governance enhances both efficiency and reviewer agreement. The HITL workflow therefore meets the intent of 21 CFR Part 11 and Annex 11 for transparent, traceable decision-making, demonstrating that explainable

AI and structured human accountability can coexist as regulator-acceptable mechanisms for responsible automation in GMP validation environments [7].

4.2 Ablation Analysis

To isolate the contribution of individual explainability features, we conducted ablation studies in which the rationale cards, confidence bands, and human-review checkpoints were selectively disabled.

Removing rationale cards decreased SME acceptance of AI suggestions by 11 percentage points, indicating that explicit reasoning visibility directly enhances reviewer confidence.

Omitting confidence bands increased override frequency by 7 percentage points, suggesting that quantified uncertainty helps SMEs calibrate decisions.

Operating the system in AI-only (no human-review) mode raised rework rates by 9 percentage points and lowered κ to 0.74, confirming that automation without human oversight introduces avoidable inconsistency [8].

These findings empirically support regulatory expectations that AI be implemented as a supporting technology rather than a decision-maker.

4.3 Case Study A—URS-to-Test Mapping

A representative URS package containing 180 requirements was analyzed. The NLP engine proposed three candidate test cases per requirement; SMEs accepted the top-ranked recommendation in 63% of cases, edited 28%, and rejected 9%. Mean authoring time per URS set decreased from 10.1 to 6.0 h (−41%), equivalent to a four-hour saving per cycle. Qualitative feedback from reviewers cited the clarity of rationale cards and the automatic linkage of URS → test pairs as primary contributors to productivity gains. QA inspectors noted that evidence-pack completeness improved, as rationale cards and SME comments were automatically incorporated into the audit trail.

4.4 Case Study B—Deviation Classification

Baseline human-only agreement with QA consensus for deviation-severity ratings averaged 82% accuracy (κ = 0.71). With the HITL classifier and SME review, accuracy rose to 90% (κ = 0.86), and high-impact misclassifications (major → minor or *vice versa*) declined by 24%. Reviewers attributed this improvement to the classifier's ability to surface comparable historical deviations along with SHAP-based reasoning explanations, allowing SMEs to justify or override the suggestion transparently. This mirrors the explainability requirements described in recent regulatory AI governance papers.

4.5 Interpretation and Regulatory Implications

Across both pilots, efficiency gains were achieved without compromising audit traceability or human accountability. The combination of *confidence-bounded outputs* and *mandatory SME signoffs* maintained full compliance with 21 CFR Part 11 and Annex 11 provisions. Results directly support the recommendations of the EMA Reflection Paper (2024) and FDA AI/ML Action Plan (2023), which call for transparent human oversight mechanisms and periodic performance validation for AI-enabled systems. By quantifying the effects on both performance and consistency, this evaluation provides initial evidence that HITL AI can serve as a *regulator-acceptable pathway* for controlled automation in validation processes, complementing established CSA and GAMP 5 principles.

5 The Core of HITL AI Framework

To build trust in AI, regulators need more than speed or efficiency claims. They want to see how decisions are made, who reviewed them, and whether every step can be explained later. That is where a Human-in-the-Loop (HITL) approach comes in. The framework rests on three layers: AI-assisted recommendations, explainability tools, and human oversight illustrated in Fig. 1. Each layer plays a role in keeping the process transparent, traceable, and ready for audit [9].

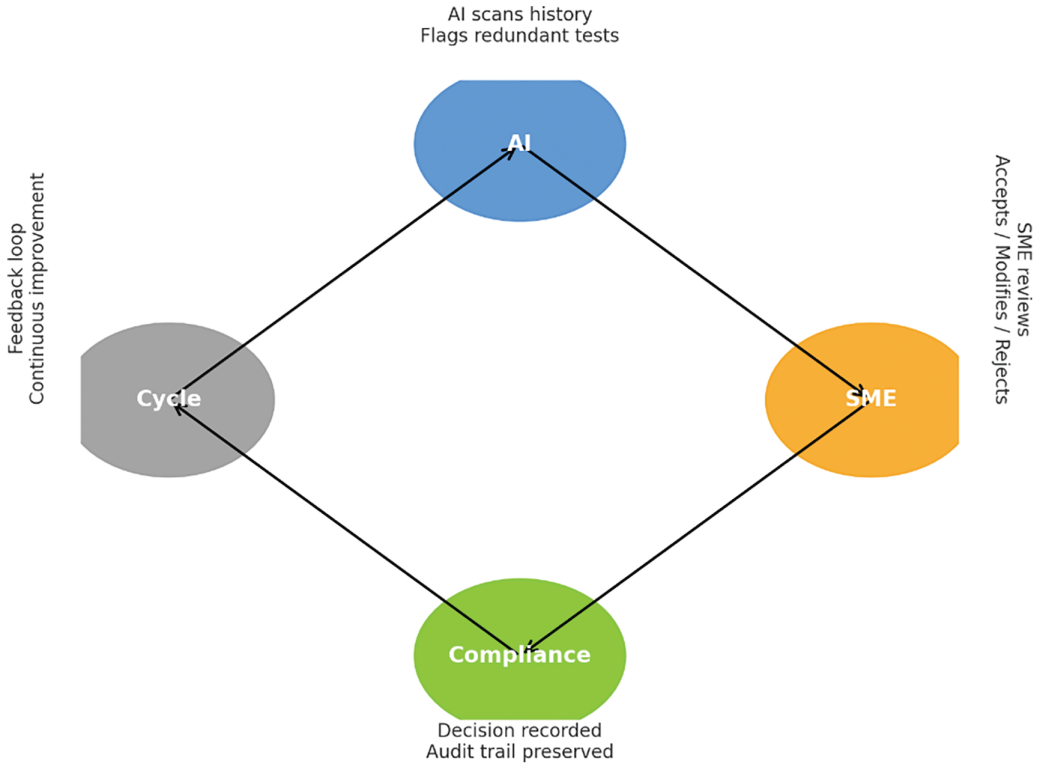

Figure 1: Framework for GMP compliance

AI-Assisted Recommendations: AI can help with the heavy lifting. Natural language processing can scan requirements and suggest test mappings. Supervised learning can help classify deviations, and analytics can point to tests that may not add much value. These are time-savers, not final decisions. Every output still flows through an SME who has the authority to confirm or reject the suggestion.

Explainability Layer: Transparency is the key to regulatory acceptance. That's why each recommendation is backed with context. Rationale cards explain why the AI proposed an option. Confidence bands show how reliable the output is. Counterfactual examples illustrate why a different path wasn't chosen. These features don't just help the SME understand the logic—they also create an audit trail that inspectors can review.

Human Oversight and Approval: In the end, people remain in charge. SMEs can accept, modify, or decline any AI recommendation. Their choices are logged with electronic signatures and time stamps, following FDA 21 CFR Part 11 and ALCOA+ principles. This preserves accountability and ensures AI is always supporting, not replacing, human judgment.

The strength of the framework lies in balance. AI speeds up routine work. Explainability makes the logic clear. Human oversight keeps responsibility in the right hands. Together, the three layers turn AI from a "black box" into a structured decision-support tool that fits within GMP expectations.

The framework comprises AI recommendations, an explainability layer, and human oversight with full auditability. Unlike prior CSA/GAMP-aligned automation tools, our framework formalizes auditable explainability artifacts (rationale cards, confidence bands, counterfactuals) and governance controls (risk tiering, challenge cadence, rollback triggers) that make AI outputs inspectable and reproducible under GMP. Table 3 further summarizes the roles and responsibilities (RACI) that govern decision rights and accountability within the HITL AI workflow.

Table 3: Roles and responsibilities (RACI)

Activity	Validation Eng.	SME	QA Manager	IT/Model owner
Review AI output	R	A	C	C
Approve/override	C	R	A	I
Challenge testing	I	C	C	R
Audit pack assembly	I	C	R	C

Approval Workflow: AI Suggestion → SME Review/Comment → QA Review (risk-tiered) → e Signature → Evidence Pack → Audit Trail. Each transition captures ownership, timestamp, and rationale within the platform, enabling complete traceability consistent with 21 CFR Part 11 and Annex 11. The corresponding diagram Fig. 2 illustrates these decision rights, sign-off checkpoints, and the governance artifacts generated at each step.

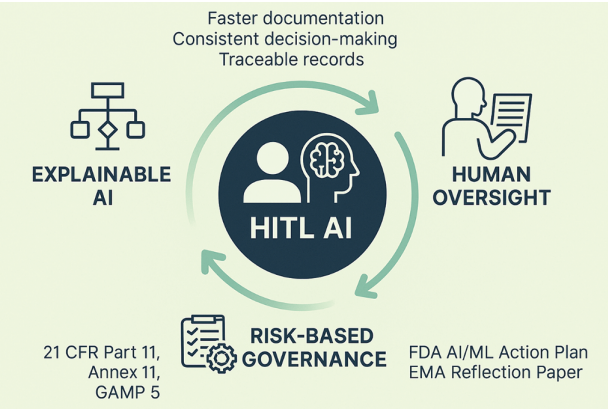

Figure 2: Approval workflow

6 Where HITL AI Adds Value in Compliance Work

The practical benefits of Human-in-the-Loop AI become clear when we look at how it fits into daily validation work. Below are three common areas where this approach can shorten timelines and improve consistency, while still ensuring subject matter experts (SMEs) remain in charge of final decisions. Three common validation activities risk-based test reduction, URS-to-test mapping, and deviation classification illustrate where this approach delivers the most value. A comparison between traditional practices and HITL AI-enabled workflows is summarized in Table 4.

Table 4: Traditional vs. HITL AI

Validation task	Traditional approach	HITL AI approach
Risk-based test reduction	Manual reviews; often too much testing just to fill the record	AI flags low-value tests; SMEs confirm or reject, aligned with CSA
URS-to-test mapping	Manual linking; inconsistent and error-prone	NLP suggests links; SMEs check and approve
Deviation classification	Human-only judgment; different reviewers give different results	AI predicts severity; SMEs finalize with context and rationale

6.1 Risk-Based Test Reduction

In many validation projects, the same problem comes up again and again too many tests. Teams keep repeating steps mainly to bulk up the paperwork. Whole weeks can be spent on activities that almost never fail, and the result is wasted time, budget overruns, and tired reviewers.

The FDA's CSA guidance, published in 2022, was a reminder to move away from that approach. It encourages companies to cut back on unnecessary repetition and to put energy into the tests that actually protect patient safety and product quality [10].

AI can support this shift. By looking at historical data, it can point out functions that have always passed without issue and suggest they may not need the same level of testing. For example, if a login check or standard system function has never failed across dozens of similar systems, the tool might highlight it for possible reduction. Even so, the subject matter expert makes the call. If there's any unique risk with that system, the test stays. The point isn't blind automation it's speeding up the work while keeping judgment and accountability in human hands.

6.2 URS-to-Test Mapping

Manually connecting user requirements to test steps is one of the slowest tasks in validation. People comb through long documents, and mistakes or inconsistencies slip in.

With HITL AI, natural language tools read the requirement and suggest possible test links. Say a URS states, "the system shall generate an audit trail." The AI might recommend a ready-made audit trail test. The SME checks the suggestion, tweaks it if needed, or discards it. The benefit is consistency across projects without losing accountability [11].

6.3 Deviation Classification

Classifying deviations is another time sink. Two reviewers might look at the same issue and land on different ratings. AI brings a starting point by comparing new deviations to past records. If it looks like previous "minor" issues, it suggests the same. The SME then decides whether that fits the current context. If they believe the risk is bigger like an impact on sterility, they can bump it up to "major" or "critical" and record the reason. This approach combines AI's speed with the SME's judgment, creating more consistent results [12].

7 Governance and Risk Controls

Bringing AI into validation is not just about clever algorithms. Without the right guardrails, even a smart system can lose reliability or drift away from what regulators expect. That's why governance is the backbone

of Human-in-the-Loop AI. It keeps the technology transparent, accountable, and inspection ready. Three practices matter most: model risk tiering, regular challenge tests, and evidence packs.

7.1 Different Models, Different Levels of Oversight

Not every AI model carries the same level of risk. A tool that suggests formatting changes in a document is low risk. A model that recommends cutting test cases from a qualification package is high risk. Tiering helps companies apply the right level of oversight.

- **Low-risk models** can be monitored lightly.
- **High-risk models**, like those used for test reduction or deviation classification, require stricter checks, validation, and SME review.

This mirrors the principles of ICH Q9 (R1) and ensures resources are spent where they matter most [13].

7.2 Periodic Model Challenges

AI is not "set and forget." Data changes. Systems evolve. What worked last year may not hold up today. That's why periodic challenge tests are critical. In practice, this means deliberately testing the AI on fresh or unusual data, then comparing its output against SME decisions. For example, a model trained on packaging deviations might fail when applied to biologics. Regular stress tests catch this drift early and keep the system reliable. Like any validation activity, results should be documented so they can be shown to inspectors [14].

7.3 Evidence Packs

Inspectors are rarely satisfied with just seeing the result. They almost always want to know how the decision was made, who signed off, and what evidence supports the outcome. That's the point of an evidence pack it brings everything together in one place. A pack typically includes:

- AI's original recommendation,
- A rationale card explaining the logic,
- The SME's decision and notes,
- A timestamped signature,
- The full audit trail.

A solid pack doesn't need to be complicated. At a minimum, it should capture the AI's recommendation, an explanation of why that suggestion was made, the SME's decision, and the signature that shows who approved it and when. The audit trail then ties it all together.

Take a simple example. Suppose the AI recommends skipping a test that has always been passed in similar systems. The rationale card explains the logic. The SME reviews it and either agrees or decides the test is still necessary. Their choice is signed, time-tamped, and stored. Later, if an inspector asks, "Why was this test dropped?", the full chain of reasoning is right there in the pack.

The benefit goes beyond compliance. Evidence packs also make life easier for SMEs and auditors. Instead of digging through scattered records or trying to recall past decisions, the story is already assembled. Inspectors get what they want, transparency and accountability, and companies avoid long back-and-forth explanations [15].

7.4 Why Governance Matters

Governance is what separates a useful automation tool from one that inspectors will not trust. Even the most sophisticated AI models can drift or yield unreliable recommendations if they operate without clear

oversight. Regulators expect evidence that every recommendation remains explainable, that review responsibilities are predefined, and that escalation occurs automatically when risk thresholds are crossed. Within the Human-in-the-Loop (HITL) framework, governance provides that structure by defining when human intervention is mandatory, how QA oversight is triggered, and how all decisions are documented. Fig. 3 depicts the overall governance and risk-control architecture that integrates model-risk tiering, challenge testing cadence, and rollback triggers.

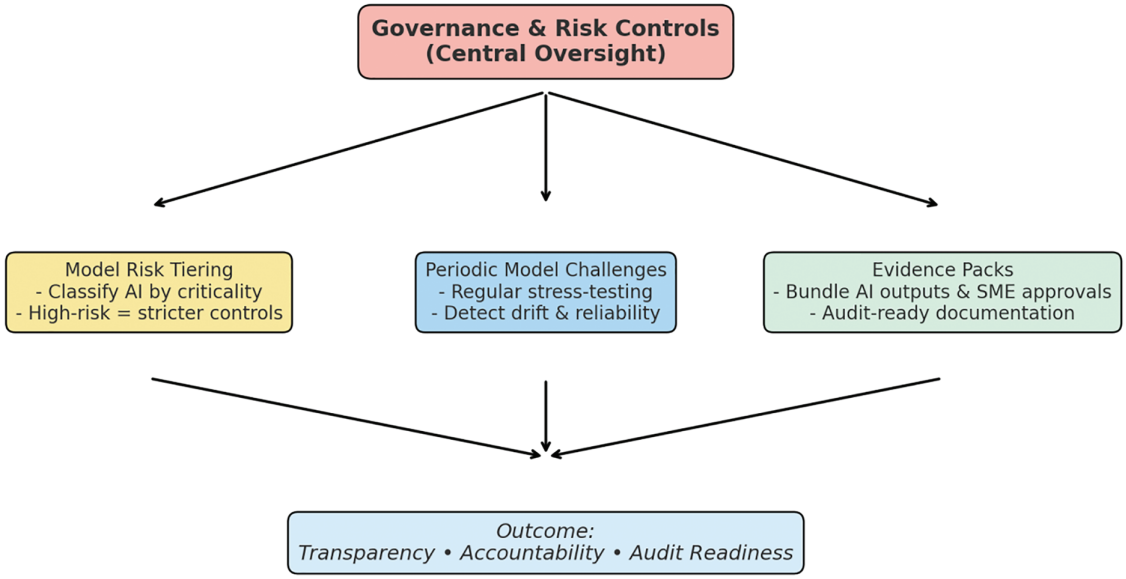

Figure 3: Governance and risk controls for HITL AI in GMP validation

To translate these concepts into operational rules, Table 5 converts the narrative controls into explicit, auditable decision logic. Each row defines the event type, the applicable model-risk tier, the confidence band threshold, the SME and QA actions required, and the corresponding regulatory clause. Together, these mappings ensure that automation remains transparent, traceable, and inspection-ready under 21 CFR Part 11, Annex 11, GAMP 5 (2nd Edition), and ICH Q9 (R1).

Table 5: Governance decision and regulatory mapping framework

Event type	Model risk tier	Confidence band	SME action	QA escalation	Regulatory clause
Test reduction > 10%	Tier 3 (High)	<0.8	Accept → Justify	Mandatory	21 CFR Part 11.10(e); Annex 11 § 12
Deviation reclassification	Tier 2	Any	Modify	Conditional (if rationale missing)	GAMP 5 § 3.3.4
Model drift > 10% in accuracy	Tier 3	N/A	Auto-rollback	Mandatory	CSA § 4.3; ICH Q9 (R1) § 5.2
AI confidence < threshold	Tier 1	<0.7	Reject	Optional	ALCOA + Integrity Principles

These structured rules operationalize the governance model shown in Fig. 3. For instance, if a Tier 3 model recommends a 15% test-reduction at 0.78 confidence, the SME must record justification, QA must counter sign before release, and the platform automatically links rationale cards, e-signatures, and timestamps to a unique evidence-pack ID (e.g., EVP-112-2025). During periodic challenge tests, if model accuracy declines by more than 10% or SME override rates exceed 20% for two consecutive cycles, the model automatically enters rollback status under change control, pending retraining and re-validation.

By formalizing these triggers and linkages, the HITL governance model ensures that every AI output remains under human supervision and fully traceable to its origin. Governance therefore acts not as a constraint but as a mechanism of trust by design, allowing organizations to embrace AI-enabled validation while demonstrating continuous control and accountability exactly what regulators expect in modern GMP environments.

7.5 Transparency, Traceability, and Data Governance

Rationale Cards: Each AI-generated recommendation is accompanied by a *rationale card* that summarizes the reasoning behind the output in a clear, auditable form. For example:

URS-112: "System shall generate Part 11-compliant audit trail".

AI Suggestion: Map to AUD-T-01; confidence = 0.91.

Key Features: "Part 11", "audit trail", "electronic records".

SME Action: *Accepted*—added comment "include archival verification step".

Trace Record: e-signature, timestamp, immutable audit-log ID A-44721.

Such rationale cards allow reviewers and inspectors to see *why* a decision was proposed, *how* it was verified, and *who* approved it, satisfying the expectations of transparency and accountability under 21 CFR Part 11 and Annex.

Traceability Matrix: Each decision is digitally linked from the original requirement through to the executed test and supporting evidence:

URS → AI suggestion(s) → SME decision → protocol section → test evidence → deviation (if any) → evidence pack ID.

This matrix ensures that every AI-assisted recommendation can be traced back to its origin and justification. It also enables one-click retrieval of all related approvals and comments, demonstrating full lifecycle control demanded by GAMP 5 (2nd Ed.) and ICH Q9 (R1).

Data-Governance Practices: All training and validation data are de-identified before model use; role-based access control (RBAC) enforces least privilege access. Data are retained according to approved SOPs and deleted or archived at the end of their defined retention period. Vendors providing AI modules undergo supplier qualification and maintain service-level agreements (SLAs) defining responsibilities for security, model updates, and incident reporting. Any modification to an AI model trigger change control and re validation to confirm continued compliance. All contributors complete conflict-of-interest (COI) declarations to ensure objectivity.

Together, these practices transform explainability features rationale cards, confidence bands, and decision logs into tangible regulatory evidence. They demonstrate that automation within the HITL framework remains transparent, traceable, and fully accountable, fulfilling the transparency and governance goals described by the FDA AI/ML Action Plan (2023) and the EMA Reflection Paper (2024).

8 Balancing Benefits with Risks in AI-Enabled Validation

8.1 Benefits

The Human-in-the-Loop AI approach offers several clear advantages for pharmaceutical validation.

8.1.1 Faster Authoring and Review Cycles

One of the clearest wins is time. AI can pull together draft mappings, spot repetitive tests, and even suggest deviation categories in seconds. What used to take hours or days for SMEs now starts with a draft they can fine-tune. In practice, this has cut validation timelines by 25%–40% in early pilots [16]. That means systems go live faster while compliance is still protected.

8.1.2 More Consistency between Reviewers

When different SMEs look at the same requirement or deviation, their answers don't always match. AI gives a baseline recommendation. Even if the SME doesn't agree, starting from the same point reduces variation. Over time, this creates more predictable outcomes across projects and sites.

8.1.3 Better Inspector Confidence

Auditors don't like uncertainty. They want to know why a decision was made and who approved it. HITL AI helps here. It produces rationale cards, confidence bands, and full audit trails that show both the AI's suggestion and the SME's final call. For inspectors, this feels like transparency instead of a "black box." That builds trust.

8.1.4 Broader AI Impact

Like healthcare more generally, where AI has already demonstrated potential to transform both clinical and administrative processes [17], HITL AI in pharmaceutical validation shows how the technology can deliver measurable efficiency gains while maintaining compliance.

8.1.5 Fits with Modern Guidance

Frameworks like FDA's CSA and ISPE GAMP 5 (2nd Edition) push for critical thinking and risk-based testing. HITL AI is naturally aligned with these expectations—it reduces unneeded work, documents SME reasoning, and focuses effort on what matters most.

8.2 Limitations

At the same time, there are challenges that must be acknowledged if HITL AI is to be adopted responsibly.

8.2.1 Need for SME Training

Not every professional validation is used to concepts like rationale cards or confidence bands. Without training, these features may confuse more than help. Companies will need to invest in educating staff so they can use AI output effectively [18].

8.2.2 Regulatory Acceptance Is Still Evolving

While regulators are open to digital tools, formal guidance on AI use in validation is thin. This means acceptance can vary. One inspector may welcome it; another may question it. Companies must be ready to explain their approach in detail until global standards catch up.

8.2.3 Extra Work for Governance and Infrastructure

Implementing HITL AI isn't a plug-and-play add-on. It needs governance, documentation, challenge testing, and integration into existing quality systems. For smaller companies, or those with limited resources, this may be a significant hurdle.

8.2.4 Risk of Over-Reliance

If SMEs begin to trust AI too much, they may stop applying the critical thinking regulators expect. This is the most subtle risk. AI is not a replacement for human judgment. Strong governance is needed to make sure people stay engaged and accountable.

The Human-in-the-Loop approach brings real benefits: faster timelines, more consistent reviews, and stronger regulatory trust. But it also comes with challenges. Success depends on training, governance, and making sure humans remain in control at every step.

9 Toward Global Harmonization of AI

Human-in-the-Loop AI is already proving its value in validation, but it's only the beginning. Over the next few years, other technologies and regulatory shifts will shape how it develops.

9.1 IoT Integration

Pharma sites are filling up with sensors—tracking temperature, humidity, pressure, and equipment performance in real time. When this data is tied to HITL AI, validation can move from periodic checks to continuous assurance. Imagine AI spotting a drift in cleanroom humidity as it happens, sending an alert for the SME to review and approve. That's proactive compliance, not reactive fixes [19].

9.2 Digital Twins for Predictive Assurance

Digital twins virtual models of equipment or processes are gaining traction. When linked with HITL AI, they could predict problems before they show up in production. For example, during a scale-up, the digital twin might simulate risks, while AI suggests which tests to run. The SME reviews, challenges, and approvals. This makes tech transfer smoother and cuts down on late surprises.

9.3 Generative AI for Protocol Authoring

Tools like large language models are already good at writing. In validation, they could be used to draft protocols, risk assessments, or test scripts. The idea isn't to hand over the process—it's to take care of the routine wording and let SMEs polish the details. Most reviewers would rather spend their time looking at real risks instead of retyping standard language [20].

9.4 Toward Global Harmonization

Right now, one of the biggest headaches is that regulators don't always agree. The FDA may be comfortable with one approach, while the EMA or PMDA may take a different view. That creates uncertainty

for companies working across regions. Over time, as more firms put HITL AI into practice and regulators compare notes, the rules are likely to get closer together. Once that happens, companies will have more confidence in rolling these practices out worldwide.

Human-in-the-Loop AI isn't the end goal, it's a steppingstone. Right now, it gives companies a way to speed up documentation and reviews without losing control. But as new tools arrive IoT sensors, digital twins, even generative AI the opportunities will expand, and it's illustrated in Fig. 4. Imagine a site where real-time sensor data feeds into validation checks, or where a digital twin of a reactor highlights risks before scale-up. These aren't futuristic ideas anymore; pilot projects are already testing them. What makes HITL AI important is that it creates a structure companies can build on. It shows regulators that AI can fit into GMP processes without breaking trust. The future will also depend on how agencies respond. Right now, FDA, EMA, and others don't always take the same position, and that makes global companies cautious. Over time, as more firms put these systems in place and regulators see the evidence, we'll likely see closer alignment. That shift will be just as important as the technology itself. At the core, one rule won't change: people stay accountable. AI can scan, sort, and suggest, but inspectors will still expect to see SME judgment on the record. For teams in the field, that means AI will handle repetitive work, while humans continue to carry responsibility. It's not about replacing expertise it's about giving experts more time to focus on risks that really matter.

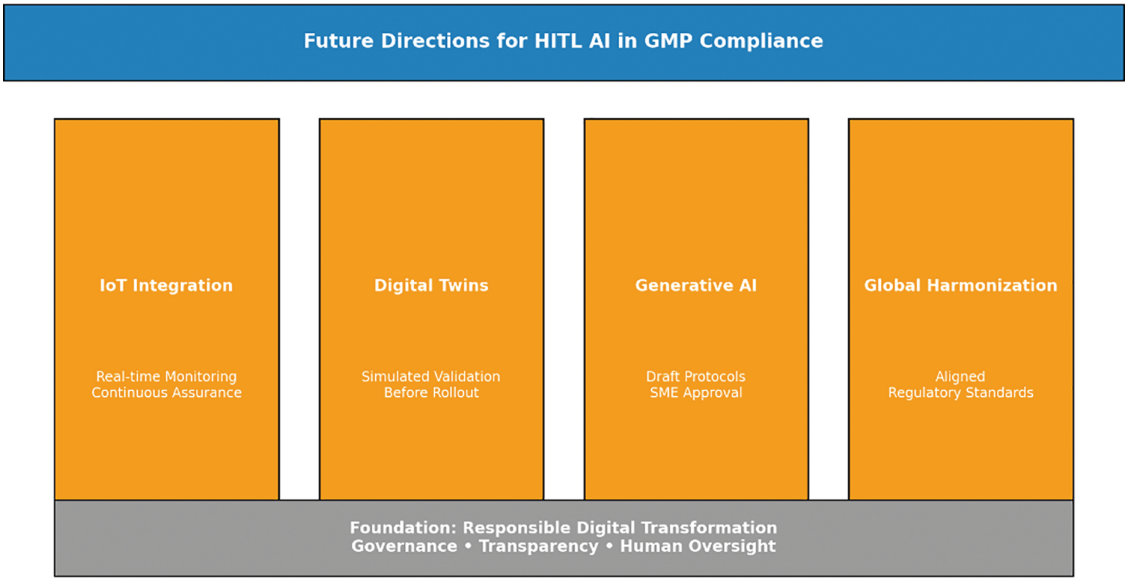

Figure 4: Pillars of future innovation in HITL AI

10 Discussion

The evaluation results show that Human-in-the-Loop (HITL) AI can significantly shorten validation timelines while improving reviewer consistency and transparency. However, its true contribution lies not only in efficiency but in how it redefines accountability within digital validation ecosystems. HITL AI shifts the emphasis from automation to augmented decision-making. By embedding explainability artifacts rationale cards, confidence bands, and counterfactual examples into auditable workflows, it aligns machine outputs with the interpretability and documentation standards mandated by CSA and GAMP 5 (2nd Edition). This creates a reproducible compliance framework where AI recommendations are visible, justifiable, and reversible.

From a regulatory standpoint, the approach resonates with recent oversight trends such as the FDA AI/ML Action Plan (2023) and the EMA Reflection Paper on AI (2024), both of which emphasize explainability, human control, and continuous model-performance monitoring. By integrating these expectations into the validation lifecycle, the proposed framework provides a transparent structure that regulators can more easily audit and trust. Beyond compliance, the findings suggest an emerging cultural shift in pharmaceutical validation: SMEs evolve from manual document authors to risk analysts and decision stewards, supported by interpretable AI tools. This redistribution of effort enables higher-value activities such as critical thinking, risk assessment, and strategic process optimization without diminishing human accountability. Nevertheless, challenges remain. The pilot data were limited to two projects within a single platform, and regulatory acceptance will depend on sustained evidence across multiple sites and products. Future research should extend these findings using standardized KPI definitions, cross-platform replication, and formal CSA-aligned model-validation protocols.

In summary, the HITL AI framework demonstrates that AI and compliance need not conflict. When explainability, traceability, and human oversight are engineered into the workflow, automation can advance both productivity and regulatory confidence, laying the foundation for trustworthy AI adoption in GxP environments worldwide.

11 Conclusion

Validation has always been about finding the right balance pushing innovation forward while keeping control firmly in place. The Human-in-the-Loop (HITL) AI framework described in this study shows that it is possible to modernize validation without weakening trust.

When explainable AI is paired with clear human oversight and practical, risk-based governance, the results speak for themselves: faster documentation cycles, fewer inconsistencies between reviewers, and a clearer record of how and why each decision was made. These improvements were achieved while still meeting the expectations of 21 CFR Part 11, Annex 11, and GAMP 5 (Second Edition).

Findings from the two pilot projects suggest that HITL AI can serve as a credible and regulator-friendly approach to introducing automation in GMP settings. As regulatory thinking continues to evolve under the FDA AI/ML Action Plan (2023) and EMA Reflection Paper (2024), the same governance tools rationale cards, evidence packs, and periodic challenge testing offer a practical template for maintaining compliance across future AI systems.

At its core, this work reinforces a simple idea: technology can enhance accountability when it is used responsibly. By keeping people at the center of the process and ensuring that every automated suggestion is transparent, traceable, and reviewable, organizations can use AI to strengthen rather than replace human judgment and build lasting confidence in digital validation.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest to report regarding the present study.

References

1. Goodman KW. Ethics, medicine, and information technology. Cambridge, UK: Cambridge University Press; 2020.
2. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58(3):82–115. doi:10.1016/j.inffus.2019.12.012.
3. U.S. Food and Drug Administration (FDA). Artificial Intelligence/Machine Learning (AI/ML) based software as a medical device action plan. Silver Spring, MD, USA: FDA; 2023.
4. Pasas-Farmer S. Governing artificial intelligence in the pharmaceutical industry. Regul Sci Innov. 2025;4(1):1–15.
5. European Medicines Agency (EMA). Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. Amsterdam, The Netherlands: EMA; 2024.
6. Higgins D, Johner C. Validation of AI-containing products across regulated healthcare industries. J Pharm Innov. 2023;18(4):112–28. doi:10.21203/rs.3.rs-2153749/v1.
7. European Medicines Agency (EMA). EudraLex volume 4, annex 11: computerised systems. Brussel, Belgium: European Commission; 2011.
8. Moreno-Sánchez P, Del Ser J, van Gils M, Hernesniemi J. A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, tradeoffs and challenges for clinical adoption. arXiv:2504.19179. 2025.
9. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923. 2017.
10. U.S. Food and Drug Administration (FDA). Computer software assurance for production and quality system software: draft guidance. Silver Spring, MD, USA: FDA; 2022.
11. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, et al. Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: ACM; 2020. p. 648–57. doi:10.1145/3351095.3375624.
12. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev. 2023;56(4):3005–54. doi:10.1007/s10462-022-10246-w.
13. International Council for Harmonisation (ICH). Q9(R1): quality risk management. Geneva, Switzerland: ICH; 2023.
14. International Society for Pharmaceutical Engineering (ISPE). GAMP 5: a risk-based approach to compliant GxP computerized systems. 2nd ed. North Bethesda, MD, USA: ISPE; 2022.
15. Yuan H, Kang L, Li Y, Fan Z. Human-in-the-loop machine learning for healthcare. Med Adv. 2024;2(3):318–22. doi:10.1002/med4.70.
16. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8. doi:10.7861/futurehosp.6-2-94.
17. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium TP. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. doi:10.1186/s12911-020-01332-6.
18. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160. doi:10.1007/s42979-021-00592-x.
19. Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7.
20. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793–813. doi:10.1109/TNNLS.2020.3027314.