SCHILLING VIGILANCE WHITE PAPER SERIES

WHITE PAPER NO. 10

# FROM AI PILOT TO GxP PRODUCTION

## A Stage-Gated Implementation Model for Pharmacovigilance

Reinhold J. Schilling

Revision 0 | 28 September 2026

Scientific and operational white paper

EU human-medicines pharmacovigilance

Evidence reviewed to 28 September 2026

schilling-vigilance.com

## Document information

| | |
|---|---|
| **Author** | Reinhold J. Schilling |
| **Series** | Schilling Vigilance White Paper Series |
| **Number** | 10 |
| **Revision** | 0 |
| **Publication date** | 28 September 2026 |
| **Scope** | EU human-medicines pharmacovigilance |
| **Evidence review date** | 28 September 2026 |
| **Website** | schilling-vigilance.com |

## Regulatory and Evidence Status

Binding EU legislation, final regulatory guidance, reflection papers and principles, consultation drafts, consensus guidance and recognised practice have different legal statuses. The Controlled Scope Envelope, Evidence Escalation Principle, Five-Stage Production Path, Six Decision Gates, Stage Regression Rule and Net Benefit Gate are author-proposed operational frameworks, not regulatory requirements. Human Correction Burden is an author-proposed operational measure. This white paper is not legal advice, a validation record, an industry standard or evidence of regulator endorsement. Applicability, including the relevant pharmacovigilance, privacy and AI-system obligations, must be determined for the particular system, operator role and documented context of use. A detailed source-status assessment appears in Appendix 1. [1–12,17]

# 1. Executive Summary

The defensible release unit is the bounded human–AI service in its documented context of use, not the model or vendor tool alone. A successful demonstration establishes possibility; production requires evidence that the service protects patients, regulated records and reporting timelines while operating reliably, remaining controllable through change and offering measurable operational value. The marketing authorisation holder (MAH) remains accountable for its pharmacovigilance (PV) processes, including those supported by suppliers. [1,2,7,8]

Model performance is only one layer. Application behaviour, workflow hand-offs, effective human oversight, end-to-end records and resilience, and production surveillance require their own evidence. This paper joins six author-proposed concepts into one decision architecture: the Controlled Scope Envelope specifies what is approved; the Evidence Escalation Principle determines depth of assurance; the Five-Stage Production Path increases operational consequence; the Six Decision Gates authorise progression; the Stage Regression Rule addresses material change or deterioration; and the Net Benefit Gate tests value after Human Correction Burden (HCB). Neither the five stages nor the six gates is legally mandated. [7,12]

The default path runs from feasibility prototype and controlled sandbox through silent mode to restricted production and scaled routine operation. Gates 1 and 2 frame experimentation; Gate 3 permits silent mode; Gate 4 permits bounded live release; G5 determines whether restricted production remains acceptable and whether preparation of a defined G6 scale proposal may begin. G6 alone authorises each specific scale increment. A justified low-risk use may combine or omit stages, but cannot bypass an applicable legal obligation or critical safety control. Silent-mode success is not proof of safe live reviewer reliance. [7,12]

### Recommendations for leaders

1. Define the task, decision owner and prohibited uses before selecting a tool.

2. Classify the actual data flow and AI use under applicable law, not by the label “PV AI.”

3. Build an adjudicated manual baseline and preserve difficult and rare cases for testing.

4. Test interfaces, misses, reviewers, records and fallback as well as the model.

5. Use silent mode where no-impact prospective evidence is informative; justify any alternative.

6. Release live use within a named, expiring and technically enforced scope.

7. Monitor HCB, reviewer detection, queue age and net benefit in normal operations.

8. Narrow, regress, suspend or retire the service when evidence no longer supports its use.

# 2. Relationship to PHAIR and White Paper No. 8

PHAIR assesses whether a defined AI use is suitable for a specific task and determines the required depth of assessment. White Paper No. 8 addresses qualification and oversight of AI-enabled PV service providers and their operational dependencies. This white paper governs how an accepted candidate progresses from controlled experimentation through bounded production to evidence-based scaling. The three are complementary: none replaces applicable validation or assurance activities, supplier qualification, privacy assessment, information-security assessment, legal review, change control or applicable quality-system processes. This positioning describes their distinct functions; it does not assert a formal cross-approval or shared document identifier.

### Related white papers in the series

• Schilling R. PHAIR: Pharmaceutical AI Risk & Readiness Framework. September 2026.

• Schilling R. Qualifying AI-Enabled Pharmacovigilance Service Providers: A Practical Due-Diligence and Oversight Framework. Schilling Vigilance White Paper No. 8, Revision 0, September 2026.

## 3. The Gap Between a Tool and a Process

> **Central proposition**
>
> A pilot shows that something can work; production evidence must show that the complete process continues to work acceptably under real operating conditions.

A locally useful drafting assistant or high-scoring extraction demonstration may still bypass case queues, leave correction effort undocumented or fail on poor-quality source material. No numerical industry-wide pilot failure rate is claimed. Scientific and regulatory sources instead support scrutiny of generalisability, verification and lifecycle controls. [7,12,20]

**Table 1. Diagnostic barriers to controlled production.**

| Barrier | Evidence gap | Appropriate response |
|---|---|---|
| **Technical** | Unstable extraction, weak languages, unreliable interfaces | Independent challenges and transfer tests |
| **Evidence** | Selected demos, leakage, rare-event gaps, disputed truth | Holdout, adjudication and prospective sampling |
| **Governance** | No owner or bounded intended use | Approved charter and decision record |
| **Integration** | Bypassed clocks, queues or regulated records | End-to-end hand-off and reconciliation |
| **Operational** | Nominal review, fatigue, no tested fallback | Reviewer-catch and timed recovery exercises |
| **Economic** | Hidden corrections and escalation | Like-for-like baseline and HCB ledger |
| **Compliance** | Unclear lawful data flow, supplier control or record reconstruction | Privacy, contract and record review |

These are diagnostic hypotheses, not a ranking of measured industry causes. The goal is proportionate evidence for a defined consequence, not a generic binder. [7,8,12]

## 4. Assess Uses, Not Tools

> **Controlled Scope Envelope, author-proposed**
>
> Tool × Version × Configuration × Data × Process × Users × Intended Use × Controls × Evidence
> The multiplication sign denotes dependence, not a numerical formula.

Configuration covers prompts,
retrieval, OCR and thresholds; evidence belongs to the population and workflow actually tested. A vendor’s qualification evidence may support local assurance, but does not approve every local configuration or PV process: the MAH must assess the use it operates. [7,8]

A tool improving wording in a human-written draft may have little direct operational consequence. The same tool selecting source facts for an individual case safety report (ICSR) can affect regulated records. Its autonomous suppression of potentially relevant safety information without effective detection may be unacceptable irrespective of aggregate accuracy. Approval therefore attaches to the defined use, not the product name. [2,7,12]

## 5. Define the Context of Use

The context-of-use charter should identify: business problem and beneficiaries; PV process, users and accountable decision maker; inputs and provenance, formats, languages, countries, products and populations; output and status, permitted and prohibited downstream actions; whether a regulated record, authority-facing text or external communication is affected; required original-source review and final authority; patient-safety and regulatory effect; false-positive and false-negative harm and detectability; volume and reporting clocks; manual baseline; data, system and supplier dependencies; lawful-data and security decisions; fallback; and prespecified success, exit and stop criteria. [2,7,8,12]

Approve a sufficiently specific charter before consequential testing or use: otherwise the reference population, endpoint and residual-risk owner are indeterminate. Stage 1 can operate under a preliminary charter and safe experimental data; that is not permission for live patient-data processing. The charter and later gate records should remain linked, not copied into several inconsistent documents. [7,10]

## 6. Risk Classification and Critical Conditions

Apply the Evidence Escalation Principle, an author-proposed framework: evidence depth increases with possible patient harm, regulatory-record effect, autonomy, exposure, irreversibility, weak error detection, sensitive data, opacity, dependence on suppliers and inability to fall back. Assess false negatives separately from false positives and evaluate whether human review actually lowers risk under expected load. Determine PV-law, GDPR and EU AI Act applicability independently for the particular use. PV affiliation alone does not make an AI system high-risk under the AI Act. [2,7,10–12,21]

**Table 2. Proposed risk-assessment outcomes.**

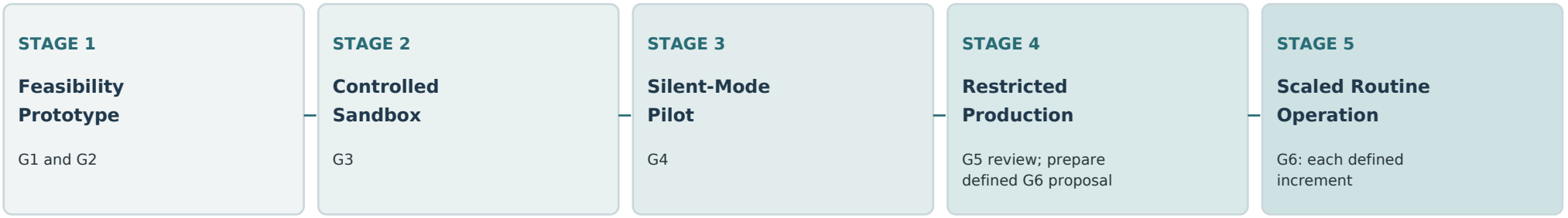
| Author-proposed outcome | Decision response |
|---|---|
| **Stop and Redesign** | Do not advance an unlawful flow or unacceptable residual safety risk |
| **Light Assessment** | Screen a bounded, non-consequential use with permitted inputs |
| **Focused Assessment** | Test reversible, source-reviewable assistance and local controls |
| **Full Assessment** | Examine critical misses, people, records and resilience where filtering, signals, deadlines or external communication matter |

Risk scores prioritise work; critical conditions determine whether a consequential step is permissible. An attractive average cannot compensate for unlawful data use, no accountable owner, missing source traceability, inability to stop, no tested fallback, undetectable material failure or unacceptable residual safety risk. These are author-proposed decision stops, not a regulator-issued checklist. A missing safeguard may permit isolated research but not the affected live use. [2,7,10,12]

## 7. The Five-Stage Production Path

The Five-Stage Production Path, an author-proposed framework, describes states of increasing consequence, not levels of model accuracy. Documented stage merging or omission can be proportionate for a genuinely low-risk use; critical conditions remain in force. Additional supervised study may be warranted for higher-consequence work. [7,12]

**Figure 1. Five-Stage Production Path and Six Decision Gates. Author-proposed framework; not a regulatory standard.**




Increasing operational consequence. Stages are not levels of model accuracy.

Stage Regression Rule: a material change or deterioration may return the affected use to an earlier stage. The response is risk-dependent and need not restart every stage.

**Table 3. Five-stage overview.**

| Stage | Purpose; permitted data/environment | Users and output | Consequence and control | Decision |
|---|---|---|---|---|
| **1. Feasibility Prototype** | Plausibility; synthetic, anonymised or otherwise approved experimental data, isolated | Builders/PV expert; research artefact | No official effect; concept review | G1/G2 |
| **2. Controlled Sandbox** | Independent test; approved historical data, segregated | Testers/adjudicators; test artefact | No official effect; reference and challenge | G3 |
| **3. Silent-Mode Pilot** | Prospective no-impact test; authorised inputs, parallel path | Shadow team; protected evaluation output | Official process unchanged; independent decisions | G4 |
| **4. Restricted Production** | Bounded live proof; eligible data, assured production | Named trained users; provisional output | Real, limited effect; source review and stop | G5 review; prepare defined G6 proposal |
| **5. Scaled Routine Operation** | Sustained wider use; explicitly approved expansion | Wider trained roster; governed records | Wider but bounded effect; monitored oversight | G6 per increment |

## 8. Stage 1: Feasibility Prototype

Objective and entry: test plausibility cheaply after naming a sponsor, preliminary task and initial privacy/security screen. Builders and a PV subject-matter expert may experiment in an isolated environment with synthetic, genuinely anonymised, licensed or otherwise explicitly approved data. Pseudonymised patient information remains personal data. Prohibit unapproved identifiable/live data, production integration, regulated records, autonomous decisions and authority-facing output. [7,10]

Keep a lightweight hypothesis and experiment log: source permission, versions, early errors, manual comparator, vendor data terms and disposal decision. Exit: a technically plausible task, initial missed-event hazards, rough correction-cost estimate, accountable process owner and credible route to representative lawful evaluation. Stop/No-Go: an unlawful or unsafe use architecture or unacceptable provider data reuse. The sponsor/PV owner decides with privacy and security input where relevant. Fallback: end the experiment and dispose of or archive material under the approved classification; the official process has not changed.

## 9. Stage 2: Controlled Sandbox

The sandbox fixes the candidate and challenges it without operational impact. After G1/G2, use approved production-like historical cases with a separate development/independent-test split, overlap checks and source-based adjudication by qualified PV reviewers. Include rare serious cases, ambiguous narratives, poor scans, local languages, negatives and out-of-scope examples. An enriched challenge set tests robustness, not natural prevalence. [7,12]

Version the model, prompt, retrieval, OCR, terminology and rules. Test source spans, abstention, role permissions, security/privacy, failed transfers, audit history, no unintended production write, reviewer usability and dry-run manual recovery. Deterministic interlocks around a probabilistic model do not make the model deterministic. Record defects, segments and uncertainty, not just pooled accuracy. [7,8,12]

G3 evidence: charter and risk record; data inventory; independent reference and disagreement log; severity-specific results and critical-miss response; supplier assessment; tested shadow isolation, reconstruction and fallback. Stop/return: leaked data, invalid test separation, unversioned update or uncontrolled critical miss. The PV process owner authorises only a protected silent-mode run after applicable Quality, IT/validation, privacy and security review.

## 10. Stage 3: Silent-Mode Pilot

In silent mode, the candidate receives prospective real-world inputs in parallel, but its output neither determines nor contaminates the official result. The existing process continues; versions, exclusions, timestamps and discrepancies are controlled. Shadow outputs are governed evaluation records, not unregulated scratch work. A protocol can blind live operators, isolate shadow output, compare it with independently produced official decisions and adjudicate disagreements against original sources. The manual decision is not automatically perfect ground truth. [7,10,12]

Silent mode can estimate prospective case-mix performance, latency, subgroup limitations, missed events and projected correction work. A separate unblinded reviewer simulation is needed to study source checking, seeded-error detection, incorrect overrides and usability. Silent mode cannot establish safe live reliance, future scale capacity, actual live fallback resilience or autonomous-decision safety. [7,12,20]

Entry: G3 and verified no-impact routing. Exit/G4: acceptable prespecified, risk-derived evidence in the approved segments, unresolved misses adjudicated, tested reviewer effectiveness and fallback, and a credible value forecast. No-Go: insufficient reference or segment evidence. Stop: shadow output affects official work, serious undetected failure, data/privacy breach or uncontrolled version change. Rollback: disable shadow and inspect any affected cases while manual processing continues. Silent mode is a useful method, not a statutory stage or an automatic prerequisite for low-risk drafting.

## 11. Stage 4: Restricted Production

Restricted production is a formal, time-limited live release. The Controlled Scope Envelope must name products, countries, languages, source types, authorised users, volume, operating hours, exclusions and approval expiry. Out-of-envelope inputs take the manual route; a convenient configuration switch cannot expand authorisation. Release requires risk acceptance, trained named staff, source-linked provisional output, abstention, effective override/stop and a staffed fallback. [2,7,8]

Run an exception queue, enhanced checks of accepted and overridden outputs, false-negative surveillance, queue-age alerts, logs, supplier support and incident escalation. Rehearse a timed manual re-route that accounts for every in-flight case. Conditional Go can cover a bounded non-critical defect with owner, due date and enhanced monitoring; it cannot cure a failed critical condition. Stop: uncontrolled official output, critical miss, lost record history, breached scope, unsustainable queue or failed fallback. Record PV owner and Quality decisions; involve the QPPV for material effects on the PV system or patient-safety oversight. [1,2,8]

## 12. Stage 5: Scaled Routine Operation

Scale requires sustained live evidence and an explicit G6 decision for each material increment of country, language, product, source, staff or volume. Reassess local reference cases, reporting effects, provider and infrastructure capacity, drift, reviewer fatigue, deputies, manual competence, disaster recovery, vendor exit and historical reconstruction. Do not replace material human checks with sampling merely because throughput increased; justify the oversight design against observed residual risk. [2,7,12]

Monitor each segment and retract an increment if its evidence fails. Persistent safety deterioration, unexplained overrides, supplier change, queue growth, lost records or negative value may require narrower scope, regression or retirement. Routine operation is an assurance obligation, not an indefinite approval. [2,7,12]

## 13. The Six Decision Gates

The Six Decision Gates, author-proposed and not regulatory gates, use one standard record: decision; exact use, version and envelope; accountable approver and independent reviewers; linked evidence; accepted residual risk and separately stated missing evidence; conditions, owner and due date; expiry and reassessment trigger; and fallback. Go permits only the recorded next activity; Conditional Go requires an enforceable safe restriction; No-Go prevents advancement. A serious live failure triggers suspension, not an administrative wait for the next meeting. [7,8,12]

**Table 4. Six Decision Gates. Author-proposed; not regulatory gates.**

| Gate and PV example | Purpose, owner and evidence | Critical / remediable question | Decision, record and response to change |
|---|---|---|---|
| **G1 Intended Use Approved; ICSR fields** | PV owner approves bounded charter, baseline and prohibited uses | Is suggestion distinct from submission? No owner/scope blocks; UI wording can be refined in research | Go to defined testing; Conditional Go only for isolated exploration; No-Go otherwise. Signed charter; revisit changed task; revert manual |
| **G2 Risks and Non-Negotiable Controls Defined; literature** | PV owner with Quality/domain review selects assurance depth; hazard, lawful-data, supplier and fallback plan | Can missed reports be found? Unlawful input/undetectable suppression blocks; noncritical test enrichment may follow | Go to controlled assessment; Conditional Go to safer synthetic sandbox; No-Go on unsafe design. Record conditions; stop affected flow on new source risk |
| **G3 End-to-End Evidence Available; duplicates** | PV owner, after validation/Quality review, authorises silent mode; independent test, versions, isolation and record proof | Can a candidate pair alter the case database? Failed isolation blocks; minor UI fix may be bounded | Go to silent mode; Conditional Go to smaller protected cohort; No-Go on leakage. Expire approval on material change; disable shadow |

## Six Decision Gates, continued

**Table 4. Six Decision Gates. Author-proposed; not regulatory gates. Continued.**

| Gate and PV example | Purpose, owner and evidence | Critical / remediable question | Decision, record and response to change |
|---|---|---|---|
| **G4 Silent Mode Passed; incoming ICSR** | PV owner/Quality, QPPV if material, authorises restricted production; prospective misses, adjudication, reviewer test, fallback and value forecast | Will reviewers catch serious omissions? Uncontained miss/no fallback blocks; exclude untested language | Go to expiring restricted production; Conditional Go narrower; No-Go remains manual/silent. Release record; recheck drift; manual re-route |
| **G5 Restricted Release Accepted; coding** | PV owner/Quality renews, narrows or suspends restricted production and may permit preparation of a defined G6 proposal; live catches, records, queues and HCB | Are important miscodes caught on ordinary shifts? Deterioration or missing records blocks; minor support issue may be time-bound | Go to continue; prepare defined G6 proposal; Conditional Go to short restricted renewal; No-Go suspends or regresses. Review expiry; restore prior safe route |
| **G6 Scale Decision Based on Monitored Performance; local journal** | Head of PV/business sponsor and PV owner, Quality review, QPPV if material; specific new-segment tests, capacity, supplier and value | Has the new language/source been evaluated? Unsafe or untested material expansion blocks; phase users if justified | Go to named scale increment; Conditional Go to smaller expiring increment; No-Go remains restricted. Reassess each increment; retract or return silent |

G1 approves bounded intended use; G2 defines risks and non-negotiable controls; G3 ends the sandbox and authorises silent mode; G4 requires silent-mode and reviewer-effectiveness evidence before authorising restricted production; G5 determines whether restricted production remains acceptable and whether preparation of a defined G6 scale proposal may begin. G6 alone authorises each specific scale increment. A remediable gap has a documented, safe closure path; accepted residual risk remains after functioning controls; unresolved uncertainty is missing evidence, not something accepted by signature.

## 14. Stage-Gate Matrix

Landscape-layout matrix. P = feasibility prototype; B = controlled sandbox; M = silent mode; R = restricted production; S = scaled routine operation. Cells are concise author-proposed controls, not universal legal requirements.

**Table 5. Stage-Gate Matrix. Author-proposed controls.**

**P = prototype; B = sandbox; M = silent mode; R = restricted production; S = scaled routine operation.**

| Dimension | P | B | M | R | S |
|---|---|---|---|---|---|
| **Purpose** | Test plausibility | Independently test candidate | Observe prospective no-impact use | Establish bounded live performance | Sustain wider approved use |
| **Permitted data** | Synthetic, anonymised or otherwise approved experimental data | Approved historical data | Authorised prospective inputs | Eligible live data | Approved expanded live data |
| **Prohibited data** | Unapproved patient/live data | Unauthorised copies | Any pathway giving AI official effect | Inputs outside approved envelope | Unevaluated new scope |
| **Users** | Builders/PV expert | Test and adjudication team | Shadow evaluation team | Named trained reviewers | Expanded trained roster |
| **Environment** | Isolated experiment | Segregated test | Isolated parallel path | Assured production | Resilient production |
| **Output status** | Research artefact | Test artefact | Protected evaluation output | Provisional until approved | Governed record with lineage |
| **Operational consequence** | No official effect | No official effect | No AI-determined effect | Bounded real effect | Wider, still bounded effect |

# Stage-Gate Matrix, continued

**Table 5. Stage-Gate Matrix. Author-proposed controls. Continued.**

**P = prototype; B = sandbox; M = silent mode; R = restricted production; S = scaled routine operation.**

| Dimension | P | B | M | R | S |
|---|---|---|---|---|---|
| **Human control** | Expert concept check | Independent adjudication | Manual decision; separate reviewer test | Source review; override and stop | Evidence-based oversight and QC |
| **Documentation/testing** | Data screen; concept test | Charter, risks; critical-case tests | Protocol; no-impact tests | Release, training; live tests | Periodic file; recovery/regression |
| **Performance evidence** | Early manual comparator | Independent reference and uncertainty | Case mix, misses, burden | Live catches, queues, HCB | Segmented trends and net value |
| **Privacy/security** | Data-use screen | Lawful basis and access tests | Shadow access and retention | Production permissions | Periodic access and retention |
| **Supplier evidence** | Data terms | Versions, updates, subprocessors | Change notice and support | Service levels and audit evidence | Capacity, continuity and exit |
| **Audit trail** | Experiment log | Source and test lineage | Inputs, output, timestamps, version | Edits, sign-off and transfer | Historical reconstruction |

# Stage-Gate Matrix, continued

**Table 5. Stage-Gate Matrix. Author-proposed controls. Continued.**

**P = prototype; B = sandbox; M = silent mode; R = restricted production; S = scaled routine operation.**

| Dimension | P | B | M | R | S |
|---|---|---|---|---|---|
| **Change control** | Log experiments | Fix tested baseline | Freeze or reset evaluation | Approved change and regression | Lifecycle and scope review |
| **Monitoring/incident response** | Feasibility; contain copies | Defects and breach route | Discrepancies; isolate shadow leak | Heightened monitoring; case impact | Segment and fatigue trends; CAPA |
| **Fallback** | End experiment | Revert test setup | Existing manual process continues | Timed manual re-route | Recovery and vendor exit |
| **Gate** | G1/G2 | G3 | G4 | G5 review; prepare defined G6 proposal | G6 authorises each defined increment |
| **Progression** | Lawful, testable task | End-to-end evidence | Prospective evidence plus oversight test | Sustained live safety; prepare G6 proposal | Evidence for next increment |
| **Regression** | Infeasible task | Invalid test/reference | Shadow contamination or critical miss | Safety, queue or record failure | Drift, supplier or scope change |

## 15. Testing the Complete Human–AI Service

Test six non-interchangeable layers: model (outputs against an independent reference); application (input, prompt, retrieval, rules and abstention); workflow (roles, queues, reporting clocks, interface reconciliation); human–AI interaction (actual error detection and correction); end-to-end service (supplier, records and continuity); and production surveillance (drift, fatigue and control performance). A successful model test validates none of the other layers. [7,8,12]

Cover functions and prohibited actions; negative and rare critical cases; poor source quality and subgroup differences; privacy/security and roles; failed and duplicated transfers; audit history and historic reconstruction; prompt/model/threshold regression; load; usability; reviewer effectiveness; fallback and business continuity. Repeat affected layers after material change. Preserve the evidence showing both what failed and whether the control prevented a regulated consequence. [2,7,12,13]

## 16. Performance Evidence and Acceptance Criteria

Core progression decisions should report critical/material error rate; false-negative rate and critical-miss count; subgroup performance; reviewer detection/catch rate; processing time; backlog and queue age; HCB; and net operational benefit. Each requires a defined unit, eligible population, denominator, reference, segment and uncertainty. A single “accuracy” measure cannot tell a PV process owner whether a missed serious report will be noticed or whether a queue remains manageable. Appendix 2 retains the extended scientific and operational catalogue, including precision, specificity, calibration, abstention, override and record reconstructability. [7,12]

Prespecify acceptance by intended use, severity, manual baseline and its limitations, detectability, compensating controls, fallback capacity, empirical data and statistical uncertainty. Report critical-event counts alongside proportions; a modest sample with zero observed misses cannot prove absence of rare harm. Keep enriched challenge results separate from routine prevalence estimates. There is no universal pilot size, duration, accuracy threshold or sensitivity target. [7,12]

## 17. Human Oversight Must Be Tested

The presence of a “qualified person” is not proof of effective oversight, and the phrase risks confusion with the formal GMP Qualified Person role. Specify a qualified PV reviewer with task-specific competence, original-source access, time for independent assessment, override and stop authority, a deputy and an escalation route. Sign-off without meaningful verification can institutionalise automation bias; evidence from other clinical settings warrants caution, not direct PV effect-size claims. [12,20]

In the sandbox, use seeded errors, blinded checks, double review, adjudication and proficiency tests. In silent mode, keep official processing independent and run a separate unblinded reviewer exercise. In restricted production, audit accepted outputs as well as overrides and monitor detection across ordinary shifts; at scale, repeat proficiency and workload checks. An override may be right or wrong; adjudicate sampled overrides before treating a rate as evidence of model or reviewer quality. [7,12]

## 18. Human Correction Burden and the Net Benefit Gate

Human Correction Burden (HCB) is an author-proposed measure of effort required to make AI-assisted work safe and complete. Count original-source review, verification, edits, documentation, escalation, repeat contact, rework, retrospective case review, quality checks, deviation/CAPA work, training and operational governance. Report attributable hours per eligible unit and the distribution by source, language, shift and severity; compare with a time-matched manual baseline of comparable quality. Silent mode can forecast HCB; restricted production must measure it. [7,12]

Net Benefit Gate, author-proposed. Use a common unit and avoid double-counting overlapping tasks:

Equation 1. Net operational benefit, author-proposed.

**Net operational benefit =**

**Comparable manual baseline**

− AI processing and support

− human review and source verification

− correction, escalation and rework

− quality and governance overhead

− bounded residual-error allowance, where defensibly quantifiable

Rare severe harm must not be reduced to an unreliable monetary estimate. Economic gain cannot override a patient-safety, legal or critical data-integrity stop; unquantified risks remain visible in the gate record. If a segment causes most rework, restrict it and retest; if reviewer time dominates, redesign the work; if sustained value is negative, restriction, suspension or retirement may be justified. Technical success is not the same as operational value. [7,12]

## 19. Worked Example: AI-Assisted ICSR Field Extraction

**Illustrative author-designed scenario.**

This hypothetical example describes no actual company, product, supplier, validated implementation, study result or regulator-approved criterion. All evidence and decisions must be replaced by local, prespecified assessment.

### Charter and Controlled Scope Envelope

A PV team considers an assistant that proposes source-linked ICSR fields from incoming English-language free-text reports for one hypothetical product family and one approved intake channel. The business problem is manual transcription effort, not autonomous case acceptance. The beneficiary is the trained case processor; the output is provisional field suggestions with source spans. Permitted use is source-based confirmation or correction before a human-controlled case record is finalised. Prohibited uses include validity or seriousness decisions, report suppression, regulatory submission, patient advice and processing unsupported languages. The Controlled Scope Envelope identifies the tool/version, OCR and prompt, input source, eligible products/language, named users, permitted workflow, review interlocks, evidence and expiry. G1 approves that charter, not a generic extraction platform. [2,7,8]

### Risk and assurance choice

Principal hazards are omitted seriousness, wrong event/date, a misidentified patient or reporter, unsupported source citation, a failed transfer and a queue that masks a reporting clock. G2 selects a Full Assessment because errors can influence regulated records and timeliness. Non-negotiable controls include lawful source access, field-level traceability, review authority, stop capability, timely manual fallback and detection of material omissions. An overall model score cannot compensate for a serious missed event. No sample size or acceptance limit is implied by this example. [2,7,10,12]

### Stage 1: Feasibility Prototype

The prototype uses synthetic narratives in isolation to see whether suggested fields have plausible source spans; its output never enters official work.

### Stage 2: Controlled Sandbox

In the sandbox, approved historical reports include poor scans, ambiguous dates and separately labelled rare-seriousness challenges. Reviewers build an independent reference and adjudicate disagreements; model, OCR, prompt and interface versions are fixed for evaluation. Tests include wrong-language abstention, write prevention, reconstruction, privacy roles, seeded reviewer errors and a manual-route dry run. G3 authorises only an isolated prospective shadow path if the evidence supports that exact candidate and its unresolved limitations are explicit. [7,12]

### Stage 3 and Gate 4

Eligible new reports are processed prospectively in silent mode while staff create official cases without AI suggestions. An independent adjudicator examines discrepancies against originals; a separate unblinded exercise tests whether representative trained PV reviewers operating under realistic conditions notice plausible but wrong AI fields. The Gate 4 packet combines segment-specific critical misses, reference disagreement, no-impact proof, expected review effort, observed latency, fallback exercise and reviewer-detection evidence. Illustrative decision: if the hypothetical shadow results support only the approved English channel while the reviewer exercise exposes weak detection for poor scans, Gate 4 might authorise a Conditional Go limited to legible inputs with mandatory source checking and enhanced QC, or a No-Go if a critical omission remains uncontained. These are alternative risk-based decisions, not findings or validated limits. A successful silent run alone does not prove safe live reliance. [7,12,20]

### Stage 4 and Gate 5

Under an authorised restricted-production decision, only named trained staff see provisional suggestions. The queue routes illegible or unsupported inputs to manual entry; operators inspect the source, correct suggestions and control the final record. Live surveillance captures critical omissions and false negatives, accepted-output QC, reviewer catch rate, transfer reconciliation, processing-time distribution, aged backlog and stop/fallback events. HCB records source checking, corrections, documentation, escalation and added QC, not just typing saved. Gate 5 compares sustained normal-shift performance and full burden with the comparable manual baseline. Illustrative decision: G5 may keep restricted production acceptable and permit preparation of a defined G6 scale proposal only if safety and record controls remain effective and net benefit remains credible; otherwise narrow or suspend. G5 does not authorise scaling. [2,7,12]

### Gate 6 and regression

A proposed German-language intake channel is a new increment, not an informal setting change. Gate 6 alone authorises the defined increment after language-specific reference material, reviewer usability and capacity tests, privacy/source checks and a feasible manual route; until approved, German reports remain manual. If the provider then changes OCR behaviour without sufficient version evidence, or live checks uncover a serious omitted event that reviewers did not detect, stop the affected route, review exposed reports and clocks, restore manual processing, and return the affected configuration to the sandbox and, if prospective generalisation changed, silent mode before any restricted re-release. This is the Stage Regression Rule, not a requirement to restart every unaffected component. [7,8,12]

## 20. Documentation, Suppliers and Decision Rights

Maintain one linked controlled evidence file, increasing depth with consequence: Stage 1 charter/data screen/experiment log; Stage 2 risks, references, versions and independent tests; Stage 3 prospective protocol, isolation and adjudication; Stage 4 release, training, live surveillance and fallback; Stage 5 periodic trends, expansion, recovery and retirement. Retain regulated PV records under applicable rules and classify temporary evaluation artefacts separately. Reconstructability matters more than document volume. [2,7,8,13]

For a hosted model, map supplier and subprocessors, hosting and data transfers, exact service identifier, data retention/training terms, change notices, logs, support, security, audit access, export and exit. If exact

deterministic replay is impossible, preserve contemporaneous input, output, version/configuration, timestamps, edits and decisions, and state the limitation. The supplier can prepare evidence but cannot assume MAH accountability for the local PV process. [7,8,10,12]

The PV process owner accepts intended-use and operational decisions; Quality independently challenges evidence and may hold an unsafe release; system/model owners maintain version and technical controls; privacy, legal and information-security functions decide within their domains; senior PV leadership funds scale; and the QPPV is involved proportionately in material PV-system or safety changes. Users need a clear route to report and stop suspected critical failure. This allocation is a proposed governance design, not a new legally prescribed RACI. [1,2,8]

## 21. Change Control, Incidents and Inspection Readiness

Stage Regression Rule, author-proposed: reassess each material model, prompt, retrieval, threshold, workflow, interface, provider, data-source, language, user-group or hosting change; document affected scope, repeat tests, decision and rollback. A minor interface adjustment may require targeted reviewer testing; a new source used for filtering may require sandbox reference cases, silent-mode evidence and restricted re-release. EMA’s AI reflection paper allows flexible PV model development, including incremental learning with validation, monitoring and documentation; its pivotal-trial frozen-model discussion should not be applied wholesale to PV. [7]

**Table 6. Change and regression examples.**

| Change or failure | Typical response; regression is risk-dependent |
|---|---|
| **New model/provider or fine-tuning** | Supplier and independent regression; B→M→R if consequential |
| **Prompt/retrieval/threshold** | Diff, omission and workload challenge; B, then M if case mix or suppression changes |
| **Workflow/interface/new user group** | Transfer, records, role and reviewer tests; B→R or M if reliance changes |
| **New language/country/product/source** | Local reference, legal/process and prospective evidence; B→M→R |
| **Hosting/subprocessor** | Privacy, security, latency and continuity; targeted B/R re-release |
| **Serious miss, drift or lost history** | Suspend scope, assess exposed cases, restore manual route; return to B/M as needed |

On an incident, contain the route and preserve records; inventory affected cases, versions, sources, users and reporting clocks; conduct independent source-based review; correct records through applicable processes; consider notification duties case by case; investigate cause and CAPA; reconcile in-flight work to the tested manual path; and approve restart on fresh evidence. A defect is not automatically externally reportable, but potential safety and privacy effects must be examined. [2,7,10,13]

Before an audit or inspection, the team should be able to show the current approved envelope; lawful input and supplier routes; independent reference and critical misses; what silent mode did not prove; live reviewer detection; a reconstructed historic decision; change/condition expiry; an in-flight fallback drill; and the evidence for the last scale increment. These are practical inspection questions, not an additional legal checklist. [1,2,7,8]

## 22. Practical PV Applications and Limits

**Table 7. Illustrative PV uses and bounded fallbacks.**

| Use and provisional output | Main hazard; critical evidence | Bounded fallback |
|---|---|---|
| **ICSR extraction: source-linked fields** | Missed seriousness/date; source and reviewer challenge | Human field confirmation; manual entry |
| **Literature screening: ranking** | Missed report; exclusions and source coverage | Defined exclusion review; manual search |
| **Duplicate detection: candidate pairs** | Wrong merge; reversible, traceable pair decision | Reviewer confirms; manual pair review |
| **MedDRA coding: term suggestions** | Miscode; coding version and expert adjudication | Coder approves; manual coding |
| **Signal prioritisation: review order** | Critical low-ranked signal; independent surveillance | Preserve critical-trigger/manual review |
| **Contact-centre triage: agent prompt** | Missed AE or inappropriate advice; local-language route | Trained agent and immediate human route |
| **Aggregate report drafting: source-linked section** | Unsupported figure; source/number verification | Accountable author edits and signs |
| **PSMF discrepancy flags** | Missed supplier/site update; controlled-source reconciliation | PSMF owner reconciles manually |

These are author-designed use concepts, not approvals. For digital sources, assess the EU-effective ICH E2D(R1) and its implementation materials alongside currently applicable GVP Module VI and local obligations. Neither a strong pilot nor this white paper demonstrates that the five-stage architecture outperforms all simpler routes. Rare catastrophic misses can escape modest samples; references may be disputed; opaque suppliers may preclude exact replay; and rare severe harm cannot be reliably monetised. Apply proportionate controls without relaxing a critical safety or legal stop. [4,6,7,12]

## Appendix 1: Evidence and Regulatory Status Note

A. Primary authoritative sources. Directive 2001/83/EC, Commission Implementing Regulation (EU) No 520/2012 read with amending Regulation (EU) 2025/1466, the GDPR and the EU AI Act as amended by binding Regulation (EU) 2026/1744 form the principal legislative baseline where applicable. Current EMA GVP Modules I, III, IV and VI, and EU-adopted ICH E2D(R1), supply process-specific final/ICH guidance with different legal statuses. An instruction in guidance is not represented here as a new statutory obligation. [1–6,10,11,21]

B. Non-binding guidance and recognised practice. EMA’s final AI reflection paper and the published EMA/FDA AI principles are regulatory reflection/principles, not legislation; CIOMS Working Group XIV is consensus guidance. EMA’s inspector Q&A explains the MAH/vendor evidence distinction. MHRA data-integrity guidance is UK final guidance; NIST AI RMF is voluntary; ISPE GAMP 5 is recognised industry practice. Human-factors findings from other settings are transferred scientific evidence, not PV-specific proof. Current EU GMP Annex 11 operates within its formal GMP scope; any PV-only use of its principles is expressly an analogy. [7–9,12–16,20]

C. Draft and consultation sources. The revised EU GMP Annex 11 and proposed Annex 22 are consultation drafts, not final PV rules. FDA’s January 2025 drug/biologics AI credibility guidance is draft and not for

implementation [19]; its February 2026 computer-software-assurance guidance is final but applies to medical-device production and quality-management software, not all EU PV AI [18]. Such texts are cited for horizon scanning or explicit comparison, never as a PV validation mandate. [17–19]

D. Claims requiring cautious wording. Regulation 2025/1466 generally applies from 12 February 2026, with its Article 1 points (7) and (9) subject to an earlier application provision; the displayed 12 August 2025 consolidation of 520/2012 warns that later amendments may not appear in it. GVP Module III Rev. 2 took effect on 10 September 2026; the current index still lists Modules IV Rev. 1 and VI Rev. 2 while further revisions are planned. EU-adopted ICH E2D(R1) took effect on 18 March 2026; consult its EU implementation strategy and applicable national instructions. Under the AI Act as amended by Regulation (EU) 2026/1744, the principal high-risk provisions for Article 6(2)/Annex III systems apply from 2 December 2027 and those for Article 6(1)/Annex I systems from 2 August 2028. These are not a universal AI Act commencement date: application varies by provision, system classification and operator role. [2–6,11,21]

E. Claims amended or omitted. This paper asserts no industry-wide pilot failure percentage or universal pilot duration, sample size, sensitivity or accuracy threshold. The 2025 consolidation alone is not described as the fully operative 2026 PV text. A pivotal-trial frozen-model discussion is not imported wholesale into PV. Neither all PV AI as AI-Act high-risk nor GMP Annex 11 as automatically binding on a PV-only system is asserted; proposed Annex 22 is not final. Supplier qualification is not local use approval; a good silent-mode outcome is not a live human-oversight test; nominal sign-off is not demonstrated effective review. [2,7,8,11,17,21]

F. Author-proposed frameworks. The Controlled Scope Envelope, Evidence Escalation Principle, Five-Stage Production Path, Six Decision Gates, Stage Regression Rule and Net Benefit Gate form a connected author-proposed decision architecture. HCB is an author-proposed operational metric. None is a regulator-endorsed standard, a statutory gate or a replacement for locally required validation and quality-system decisions. [7,12]

# Appendix 2: Extended Performance and Operational Metric Catalogue

For every measure, predefine eligible population, exclusions, index time, unit, reference/adjudication, relevant segmentation (product, language, source, country, seriousness, reviewer, period) and uncertainty. Unless otherwise stated the reference is qualified, original-source PV adjudication; operational measures use controlled time, queue or transaction records. B = sandbox; M = silent mode; R = restricted production; S = scaled routine operation. Enriched critical-event sets are not prevalence samples. [7,12]

**Table 8. Extended performance and operational metric catalogue.**

| Metric | Unit; numerator / denominator or calculation; reference | First informative stage; segmentation, limitation and gaming risk |
|---|---|---|
| **Sensitivity/recall** | True flagged events / all adjudicated true events; source adjudication | B challenge, M natural. Severity/language; excluded positives and rare denominator inflate it |
| **Precision** | True positive flags / all positive AI flags; source adjudication | B, prevalence meaningful M/R. Source/time; fewer alerts improve precision while misses may rise |
| **Specificity** | Correctly rejected negatives / all adjudicated negatives | B/M. Source; many negatives conceal critical misses |
| **False negatives** | Missed true events / all adjudicated true events, plus critical count | B/M/R. Severity/source; weak reference or selection hides misses |
| **Critical/material error** | Predefined severity-class errors / all eligible outputs or decisions; adjudicated | B→S. Severity/user; relabelling and non-reporting lower the rate |
| **Subgroup performance** | Base metric numerator/denominator within each prespecified stratum | B/M/S. Language/product/source; pooled results hide thin or untested cells |
| **Calibration** | Observed event fraction within probability bins versus stated predicted probability; adjudicated | B/M, only for meaningful probability scores. Source/time; rescaling hides poor discrimination |
| **Abstention** | Manual/no-answer routes / eligible cases | B→S. Source/language; excessive abstention may erase utility |
| **Out-of-distribution detection** | Correctly flagged unsupported inputs / adjudicated unsupported inputs; also false flags / supported inputs | B challenge, M/S. New source; known challenge cases miss unknown unknowns |

# Extended Metric Catalogue, continued

**Table 8. Extended performance and operational metric catalogue. Continued.**

| Metric | Unit; numerator / denominator or calculation; reference | First informative stage; segmentation, limitation and gaming risk |
|---|---|---|
| **Override** | Human-changed suggestions / reviewed suggestions; sample changes for correctness | B simulation, R/S actual. Reviewer/shift; discouraged edits suppress rate |
| **Correction** | Outputs with material edits / outputs reviewed; compare edits to source | B/M forecast, R/S actual. Error/source; undocumented work understates burden |
| **Reviewer detection/catch rate** | Material planted/adjudicated AI errors correctly caught / material errors exposed to reviewers | B simulation, M separate exercise, R/S QC. Shift/severity; staged artificial tests may overstate real detection |
| **Reviewer disagreement** | Disputed decisions / independently double-reviewed decisions, followed by source adjudication | B/M/R. Reviewer/complexity; agreement is not truth |
| **Review time** | Active source-check/approval minutes / eligible reviewed units; observed time record | B simulation, R/S actual. Shift/source; omitted retrieval/interruption understates effort |
| **Correction time** | Active edit/recheck minutes / corrected units; observed time record | B simulation, R/S actual. Severity; moved rework hides costs |
| **Backlog and queue age** | Open eligible cases at cut-off and their age bands / cohort or work queue; queue record | M projected, R/S actual. Deadline/source; total count conceals late tails |
| **Throughput** | Completed eligible units / staffed hours, adjusted for mix; operation record | M forecast, R/S actual. Shift; easy-case selection inflates it |
| **Processing time** | Receipt-to-completion interval per case; distribution and deadline strata; timestamp record | M parallel estimate, R/S actual. Seriousness; mean hides late critical cases |

# Extended Metric Catalogue, continued

**Table 8. Extended performance and operational metric catalogue. Continued.**

| Metric | Unit; numerator / denominator or calculation; reference | First informative stage; segmentation, limitation and gaming risk |
|---|---|---|
| **Escalation rate** | Referred items / eligible items; documented referral | M/R/S. Severity/source; withholding escalation games the rate |
| **Availability** | Available scheduled service time / total scheduled service time; service log | R/S. Site/shift; uptime omits unsafe degraded outputs |
| **Failed transfer** | Incorrect or unacknowledged transactions / all attempted transfers; reconciliation reference | B→S. Interface; only counting visible failures misses silent loss |
| **Reproducibility** | Same-version repeats within prespecified behavioural tolerance / tested repeats; fixed test set | B/R. Configuration; insisting on identical wording misrepresents stochastic behaviour |
| **Reconstructability** | Material decisions with retrievable source, versions, output and review history / sampled material decisions; historic audit | B test, R/S samples. Vendor/version; plausible replay cannot replace contemporaneous record |
| **Compliant adoption** | Eligible uses following approved workflow / eligible staffed opportunities; audit and roster | R/S. User/shift; high usage alone does not imply safety |
| **Human Correction Burden** | Total attributable review, verification, correction, escalation and rework hours / eligible units; effort ledger | M forecast, R/S actual. Segment/shift; omitted QC/CAPA creates false savings |
| **Net operational benefit** | Comparable manual resource use minus total AI-supported resources, governance and defensibly bounded residual-error allowance per unit/period | M scenario, R/S decision. Mix/time; implausible rare-harm pricing overstates gain |

## Sources

Only works cited above are listed. Status descriptions apply to the sources, not the proposed operational architecture.


1. Directive 2001/83/EC, especially Article 104 — EU directive, as transposed.
2. Commission Implementing Regulation (EU) No 520/2012, displayed consolidation of 12 August 2025 — binding, to be read with [3].
3. Commission Implementing Regulation (EU) 2025/1466 — binding amendment.
4. EMA, Good Pharmacovigilance Practices (GVP) index — final module versions and update status.
5. EMA, GVP Module III, Revision 2 — final, effective 10 September 2026.
6. EMA, ICH E2D(R1) and EU implementation information — EU-adopted ICH guidance.
7. EMA, Reflection paper on AI in the medicinal product lifecycle — final non-binding reflection paper, particularly §2.3.7.
8. EMA, Coordination of pharmacovigilance inspections: electronic-system Q&A — regulator interpretive Q&A on MAH/vendor responsibility.
9. EMA/FDA, Guiding Principles of Good AI Practice in Drug Development — January 2026 joint non-binding principles.
10. Regulation (EU) 2016/679 (GDPR) — binding EU law.
11. Regulation (EU) 2024/1689 (EU AI Act) — binding EU law as amended; read with [21].
12. CIOMS Working Group XIV, Artificial intelligence in pharmacovigilance (2025) — non-binding consensus report.
13. MHRA, GxP Data Integrity Guidance and Definitions — final UK guidance, applied to this EU discussion by explicit transfer where appropriate.
14. NIST, AI Risk Management Framework 1.0 — voluntary framework.
15. ISPE, GAMP 5, second edition — recognised industry practice, not PV law.
16. EU GMP Annex 11 (2011) — within GMP scope; PV-only principles used by analogy.
17. European Commission, consultation on revised GMP Annex 11 and proposed Annex 22 — consultation drafts, not final guidance.
18. FDA, Computer Software Assurance for Production and Quality Management System Software — February 2026 final guidance for medical-device production/QMS software; comparison only.
19. FDA, Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products — January 2025 draft guidance, not for implementation.
20. Goddard K, Roudsari A, Wyatt JC. “Automation bias: a systematic review of frequency, effect mediators, and mitigators.” JAMIA (2012) — peer-reviewed non-PV human-factors evidence, transferred cautiously.
21. Regulation (EU) 2026/1744 amending the EU AI Act — binding amendment; see also European Commission, AI Omnibus implementation notice for the simplified timetable.