Schilling Vigilance White Paper Series

White Paper No. 6 · Rev. 0 · 25 September 2026

Beyond “Human in the Loop”

Designing Effective Human Oversight in Pharmacovigilance AI

Scope: EU post-authorisation pharmacovigilance for human medicinal products.
Evidence cut-off: 23 September 2026.
Status: Professional white paper. The frameworks and measures marked *proposed* below are original analytical contributions, not regulatory requirements.

Author
Reinhold J. Schilling

Executive summary

A reviewer’s name on an AI-generated pharmacovigilance (PV) record does not establish an effective control. The person may be unable to see an omitted fact, distinguish source evidence from generated text, act before a reporting deadline, or stop an unsafe process. The human can also mistakenly override a correct result. Effective oversight must therefore be assessed at the level of the *combined human–AI workflow*: important errors detected and corrected in time, important errors missed, new errors introduced, and the resources consumed. [1–4]

EU PV law requires an effective quality system, competent resources, verifiable information, traceable records and timely reporting. GVP elaborates critical PV processes and quality-system implementation. There is no universal rule in the cited sources that a human individually inspect every AI-generated field. There **are** task-specific expectations: the current MedDRA term-selection guide calls for qualified review of term selection and human oversight of autoencoding. The final CIOMS Working Group XIV report advocates risk-based oversight and evaluation of the combined system, but does not prescribe the control modes in this paper. [1–3, 5, 6]

This white paper proposes a reviewer evidence package; a decision framework for full, sampled and exception review; Human Correction Burden (HCB); direct testing of human error detection; and a six-stage oversight maturity model. The evidence supports these as plausible methods of assurance, not as a proven universally optimal operating model. A deployment earns a claim of effectiveness only through use-case-specific evidence. [3, 4, 7]

1. What “oversight” must control

A useful control question is: **Which consequential error can occur, will the reviewer encounter it, can the reviewer recognise it from the source, and can the reviewer intervene before it matters?** “Human in the loop” usually denotes participation before a decision takes effect; “human on the loop” describes supervision of operation that may produce outcomes without

individual pre-release intervention. Neither label tells us whether the control works. Map the actual production path, including AI-negative and silently excluded records, rather than relying on the vendor's workflow label. [3, 4]

Regulatory and evidentiary boundary

Commission Implementing Regulation (EU) No 520/2012 sets requirements for the MAH's PV quality system, resources, record management and reporting. GVP Module I identifies ICSR processing and signal management as critical processes; Modules VI and IX elaborate case handling, literature and digital reports, and signal management. An AI supplier or delegated processor does not remove the MAH's accountability. A general quality-system obligation is not, however, a prescribed AI reviewer-to-case ratio, HCB target, sampling percentage or universal 100% review rule. GVP distinguishes cited legal duties, generally expressed as "shall", from implementation guidance, generally expressed as "should". [1, 2, 8, 9]

EMA's final 2024 reflection paper addresses AI across the medicinal-product lifecycle. Its close-supervision statement about AI-generated *product-information text* must not be recast as a command for universal review of all PV AI outputs. CIOMS Working Group XIV's final 2025 report provides PV-specific expert consensus on risk-sensitive human oversight and combined-system quality assurance; it is not binding legislation. The ICH-endorsed March 2026 MedDRA *Term Selection: Points to Consider* contains a narrower, specific call for qualified review of term selection and human oversight of autoencoding. [3, 5, 6]

The EU AI Act requires a system-specific Article 6 classification and applicability analysis before its Article 14 high-risk human-oversight provisions are invoked. Following the 2026 amendment, relevant high-risk provisions are scheduled for 2 December 2027 for Annex III systems and 2 August 2028 for specified Annex I product-related systems. PV use alone does not make every tool high-risk. GDPR Article 22, GMP Annex 11 and US electronic-record provisions also require their own scope tests; none is a generic substitute for applicable PV law. [10, 11]

Source hierarchy: applicable legislation is binding; GVP is regulatory guidance incorporating citations to law; EMA's reflection paper contains regulatory scientific considerations; CIOMS provides expert consensus; scientific studies supply context-dependent evidence; ISO/NIST materials are standards or voluntary frameworks; and the controls expressly described as *proposed* here are authorial designs.

2. Give the reviewer an effective evidence package

The following four-layer specification is **proposed**, drawing on PV traceability, terminology guidance and assurance principles. No single source prescribes every display or metadata field. [2–6]

Layer	What the reviewer needs	Failure if absent
Original evidence	Full original contact, article or permissible call record; received time, language, source type and location of each extracted claim.	A generated summary omits a hospital admission or a second adverse event.

Layer	What the reviewer needs	Failure if absent
Decision comparison	AI proposal beside the source and final disposition; missing, contradictory and inferred facts separated; current rule and terminology version.	A confident MedDRA code masks the reporter's less-specific verbatim wording.
Limits and consequence	Intended-use boundaries, uncertainty, known failure modes, deadline and consequence of an incorrect acceptance.	"High confidence" conceals a serious-case rule or an out-of-scope input.
Intervention	Practical accept, correct, defer, reject, escalate and stop/fallback routes; appropriately qualified reviewer and protected capacity.	An escalation button exists, but cannot prevent unsafe closure or transmission.

Source-linked highlights can accelerate verification but also anchor the reviewer on what AI chose to show. For consequential decisions, prospectively compare source-first and AI-first displays against known omissions, measuring critical-error detection, mistaken overrides and time. Retain the whole source. This is a testable recommendation, not a demonstrated PV-wide mitigation. [7, 12]

Reviewer competence must match the actual decision: a processor, medical assessor, terminology specialist or literature reviewer may be needed at different points. Check peak load, absences, shift coverage, access to original sources and manual fallback. Practical independence is particularly important for *independent QC*, but there is no proposed universal staffing or organisational-independence formula for every individual review. [1–3]

3. Choose and validate the review mode

Full review, probability-based sampling and exception review provide different protection. Full review may be shallow under overload; a queue of AI-positive exceptions cannot show what was silently excluded. An independently assessed sample of the complete disposition population—including accepted and AI-negative items—is needed to estimate operational miss rates. Enriched challenge cases can probe rare failures but cannot, without a prevalence-representative sample, estimate how frequently those failures occur in production. [3, 4]

Table 1. Proposed review-mode decision framework

Context	Proposed posture	Basis for reducing or restoring review
Initial release or material change in model, workflow, prompt, source, language or terminology	Full review of affected consequential outputs for a defined qualification period; independent challenge and subgroup tests.	Reduce only after demonstrated end-to-end performance, capacity and fallback. Re-qualify affected scope after change.
Case validity, seriousness, expedited reporting and consequential ICSR or submission content	Human determination or verification before consequential release, supplemented by risk-based independent QC.	Demonstrate the *method* finds omissions and incorrect classifications; a sign-off count does not suffice.
Automated exclusion or closure, suppressed contacts, dismissed or low-ranked potential signals	Independently check negative and non-alert populations; retain full review where missed-important-item risk is not bounded.	Demonstrate sensitivity to critical false negatives in relevant products, languages and sources.

Context	Proposed posture	Basis for reducing or restoring review
Stable, bounded, lower-consequence tasks	Consider random monitoring with risk-stratified, targeted and exception checks after prospective validation.	Predetermine performance, drift, capacity and return-to-full-review triggers.
Critical miss, unexplained drift, missing source/audit trail or failed fallback	Return affected scope to full/manual processing or suspend unsafe action.	Investigate and contain exposure; define a documented restart decision and evidence.

Before examining results, define the sampling unit and population, reference-standard adjudication, materiality, strata, inclusion probabilities, uncertainty and decision threshold. If zero misses are observed in an independent representative sample of n, an approximate one-sided 95% upper bound on the miss probability is $3/n$. Zero misses among 300 sampled decisions therefore remains compatible with roughly 1% misses. That illustration is **not** a regulatory target and is weak evidence about rare critical failures, clustered errors or unrepresented subgroups. [3, 4]

Define who can restrict a model, who can stop release, what happens to already processed records, and what test evidence permits restart. A numerical model score is not necessarily a calibrated probability or a measure of the harm of an error; a high score does not cancel a serious-case rule. [3, 4]

4. Test the human control—not just the model

Evaluate AI alone, the reviewer alone where feasible, and the complete workflow on comparable cases. The reference-standard set should include correct outputs, erroneous outputs, omitted facts, accepted errors, unflagged true positives, misleading rationales, rare languages and changed configurations. Blind independent audits must include *accepted and excluded* items. An override is not automatically correct ground truth. [3, 4, 7]

Report reviewer detection of independently confirmed critical AI errors before consequential release; false reassurance among AI-wrong outputs; incorrect human overrides; disagreement; turnaround; and performance under ordinary and peak workload. Predefine severity, adjudicate disputed reference decisions and show uncertainty by relevant subgroup. Controlled, segregated seeded-error exercises can test vigilance; never insert unlabelled errors into live regulatory submissions or transfer test risk to patients. These test designs are **proposals**, not prescribed PV qualification protocols. [3, 4]

Illustration. AI classifies an ICSR as non-serious, while the source mentions hospital admission. An approval log shows that someone accepted the field. An effective-control assessment asks whether qualified reviewers find the omitted admission before release, whether they sometimes wrongly overturn correct seriousness decisions, and how often comparable omissions escape both AI and reviewer. Those quantities require independent review of accepted records, not an audit of corrections alone.

Human Correction Burden

Human Correction Burden is an author-proposed measure of post-AI human work attributable to verification, source checking, correction, correction documentation, escalation, rework and additional QC. Allocate downstream deviation, CAPA and retrospective review to exposed cohorts where defensible, or disclose them separately. Do not mix active minutes with elapsed delay or count the same work twice. [3, 4]

$$\text{HCB per output} = \frac{\text{Total Attributable Post-AI Human Minutes}}{\text{Number of AI-Processed Outputs}}.$$

Report the distribution as well as the mean, segmented by decision consequence, version, source, product, language and workload. Pair HCB with material-correction rate, critical-error detection, false reassurance, retrospective rework and deadline breaches. An output denominator alone can hide silently excluded articles or unflagged contacts: report those exposed populations separately. Neither more corrections nor fewer corrections necessarily means better oversight without an independently assessed reference standard. [3, 4]

A broader author-proposed management calculation is:

$$\begin{aligned}\text{Net Automation Benefit} = {} & \text{Manual Baseline Effort} \\ & - \text{AI-Supported Processing Effort} \\ & - \text{Human Review Effort} \\ & - \text{Correction and Escalation Effort} \\ & - \text{Quality and Governance Overhead} \\ & - \text{Expected Cost of Residual Errors.}\end{aligned}$$

Keep all terms in consistent units and state assumptions if converting effort to money. Compare matched case mix and baseline quality; show safety and compliance outcomes separately. Rare harms, unidentified misses and uncertain counterfactuals make a single “ROI” number fragile.

5. Apply oversight to four PV pathways

The profiles below are **proposed implementations**, not universally mandated review settings. Each must be adapted to the MAH’s actual responsibilities and operating context. [1–4, 8, 9]

Pathway	Hidden failure and necessary evidence	Proposed assurance focus
ICSR processing	Omitted validity element, seriousness fact or reaction; false duplicate; inaccurate code or fabricated narrative. Retain full original contact, timeline, verbatim event, duplicate candidates and current MedDRA version.	Verify consequential decisions before release; independently audit accepted records; medical and terminology escalation; timeliness and critical-omission measures. [2, 5, 8]
Literature screening	False exclusions silently remove reports or signal evidence. Retain search/retrieval record, abstract or full text where needed, translation, selection rule and rationale.	Audit excluded strata; enrich rare positive challenge sets; avoid unvalidated autonomous exclusion by language, source or product; track false-negative exclusions and time to identification. [3, 8, 13]
Signal detection	Missed or delayed patterns and misleading rankings escape the alert queue. Retain underlying cases, deduplication, trends, relevant literature and prior clinical assessments.	Document clinical validation and dismissal; independently review lower-ranked/non-alert populations; investigate material missed patterns and stop a persistently unsafe prioritiser. [3, 9]

Pathway	Hidden failure and necessary evidence	Proposed assurance focus
Chatbot and complaint intake	Lay AE descriptions are missed, false reassurance is given, or AE-plus-complaint content fails dual routing. Preserve permitted original transcript, timestamps, language and handoff receipt.	Route potentially safety-relevant content promptly to PV and relevant quality pathways; audit negative classifications, abandoned chats and handoff failures rather than allowing unsupported autonomous closure. [2, 8]

A published literature-screening experiment used a task-specific chlorine-safety dataset and several evaluation configurations. Its numerical results cannot validate autonomous exclusion in a production medicinal-product workflow; any quoted sensitivity or specificity must be tied to the exact study sample and configuration. [13]

6. Proposed six-stage maturity model

This six-stage framework preserves the original progression from nominal human availability to demonstrated risk reduction, while incorporating the dossier's stronger requirements for risk design, independent evidence and adaptive assurance. It is **not** an EMA, CIOMS or ISO scale. [3, 4]

Level	Oversight state	Evidence needed
0 — Human available	A person is named, but no reliable encounter with consequential outputs is shown.	Map AI outputs and downstream decisions; do not call availability an effective control.
1 — Human notified	Consequential outputs or exceptions can reach an identified person within an intervention window.	Test delivery, coverage, timeliness and failed-notification escalation.
2 — Human reviews	A defined task, access to original evidence, competence and recorded disposition exist.	Reconstruct what the reviewer saw and did; assess workload and source accessibility.
3 — Human understands, challenges and corrects	Risk-specific modes, correction authority, negative-population checks and fallback function.	Prospectively test omissions, wrong advice, incorrect overrides and response time.
4 — Oversight is monitored	Independent assessments of detection, false reassurance, override quality and HCB drive action.	Demonstrate subgroup monitoring, documented triggers, reversion and CAPA-effectiveness checks.
5 — Oversight demonstrably reduces residual risk	Combined-workflow evidence shows timely detection and acceptable residual risk compared with a defensible alternative.	Show sustained, version- and subgroup-aware evidence and actual authority to restrict or stop unsafe automation.

The higher levels are conditional on continued performance. A changed source, model or workload can invalidate earlier assurance; a dashboard or approval log cannot substitute for evidence of critical-error detection.

7. Deployment conclusion

The case for human oversight is strong as a *design and assurance obligation*: PV work remains accountable, and both AI and humans can err. The case that a particular configuration provides an effective control is empirical. Ask what consequential errors the reviewer can find, which ones never appear in the queue, whether the reviewer can act in time, how often human

correction is wrong, and whether the residual risk and burden improve on a realistic alternative. [1–4]

Neither blanket full review nor sampling deserves automatic preference. Intensify review when errors are consequential, hard to detect or unbounded; test the complete workflow before reducing it; independently inspect negative populations; and maintain an operable reversion route. Direct PV-specific human-factors evidence remains limited, so results from healthcare or other safety-critical contexts should be identified as transferred mechanisms, not quoted as PV error rates. [3, 4, 7, 12]

References

These are curated principal sources, not the dossier's unfiltered URL list. Application of legal provisions and current versions should be checked for the actual system and jurisdiction.

1. European Commission. Commission Implementing Regulation (EU) No 520/2012 of 19 June 2012 on the performance of pharmacovigilance activities. EUR-Lex text.
2. European Medicines Agency and Heads of Medicines Agencies. *Guideline on good pharmacovigilance practices (GVP), Module I: Pharmacovigilance systems and their quality systems.* EMA document.
3. Council for International Organizations of Medical Sciences. *Artificial intelligence in pharmacovigilance.* CIOMS Working Group XIV report. Geneva: CIOMS; 2025. Final report.
4. *Research dossier: Beyond Human in the Loop.* Unpublished evidence and operational design brief supplied for this white paper; evidence cut-off 23 September 2026. Its proposed frameworks are distinguished here from source-based requirements.
5. European Medicines Agency. *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle.* Final; 2024. EMA document.
6. MedDRA Maintenance and Support Services Organization. *MedDRA Term Selection: Points to Consider.* ICH-endorsed guide, release 4.26; March 2026. Guide.
7. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association.* 2012;19(1):121–127. doi:10.1136/amiajnl-2011-000089. Original review.
8. European Medicines Agency and Heads of Medicines Agencies. *GVP Module VI: Collection, management and submission of reports of suspected adverse reactions to medicinal products.* Rev. Rev. 2. EMA document.
9. European Medicines Agency and Heads of Medicines Agencies. *GVP Module IX: Signal management.* Rev. 1. EMA document.
10. European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), amended by Regulation (EU) 2026/1744. Consolidated act; amendment.
11. European Commission. AI Act Service Desk: When does enforcement start? Official explanatory resource. Read alongside reference 10.

12. Wang D-Y, Ding J, Sun A-L, et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. Journal of the American Medical Informatics Association. 2023;30(10):1684–1692. doi:10.1093/jamia/ocad118. Original study. Clinical-imaging context, not PV performance.

13. Li D, Wu L, Zhang M, et al. Assessing the performance of large language models in literature screening for pharmacovigilance: a comparative study. Frontiers in Drug Safety and Regulation. 2024;4:1379260. doi:10.3389/fdsfr.2024.1379260. Original study.

Author

Reinhold J. Schilling

Correspondence

schilling-vigilance.com

Scope and limitations

This white paper addresses EU post-authorisation pharmacovigilance for human medicinal products. It presents professional analysis and proposed operational frameworks. The frameworks and measures expressly identified as proposed are original analytical contributions and are not regulatory requirements. This publication does not constitute legal advice or regulatory guidance.

Evidence cut-off: 23 September 2026.

Conflicts of interest and independence

This white paper is an independent professional publication. The views and interpretations expressed are those of the author and should not be attributed to any employer or other organisation. No external organisation influenced the analysis or conclusions presented in this paper.

Suggested citation

Schilling RJ. Beyond “Human in the Loop”: Designing Effective Human Oversight in Pharmacovigilance AI. Schilling Vigilance White Paper Series, No. 6, Rev. 0; 25 September 2026.