Paper ML05

Performance-Driven Risk Oversight for AI Agents in Clinical Research

Laxmiraju Kandikatla, Maxis AI., Edison, NJ, United States

ABSTRACT

The deployment of AI systems in healthcare demands continuous, risk-aligned oversight to ensure safe and responsible operations. We propose a performance-driven model that calibrates human involvement based on two risk dimensions: metrics risk (including accuracy, precision, recall, F1-score, transparency, bias, robustness, explainability, and usability) and AI system risk (clinical impact and decision consequence). These dimensions are integrated into a 3×3 risk matrix, where the intersection informs the appropriate level of oversight. High-risk Systems require Human-in-Command (HIC) oversight, with final decision authority. Medium-risk systems operate under Human-in-the-Loop (HITL) models, with human supervision and active feedback. Low-risk systems function under Human-on-the-Loop (HOTL) oversight, where humans monitor system outputs and intervene only when anomalies occur. For example, an AI radiology tool with high accuracy but substantial clinical consequence risk would necessitate HITL or HOTL oversight. This systematic Performance-Driven Human Oversight Framework (PD-HOF) for AI Systems balances innovation with accountability, promoting patient safety and trust in clinical research and healthcare.

1. INTRODUCTION

AI evaluation serves as a fundamental pillar in the validation and lifecycle management of artificial intelligence systems, ensuring their reliability, safety, and effectiveness across regulated and high-impact domains. Evaluation is a critical component of AI validation, providing objective evidence of system capabilities, fitness for intended use, and ongoing performance assurance, while offering confidence to stakeholders that ethical, compliance, and operational requirements are met (FDA, 2025; ISO, 2018). The primary purpose of AI evaluation is to verify that the system consistently achieves its intended use with acceptable accuracy and reliability under normal operating conditions and reasonably foreseeable scenarios (ISO/IEC, 2023; Raji et al., 2020). This requires systematic assessment of critical performance attributes, including accuracy, precision, robustness, and reliability, alongside the identification of biases, limitations, or failure modes that may compromise outputs or result in unintended harm (Mitchell et al., 2019; Raji et al., 2020). The importance of AI evaluation extends beyond technical performance metrics to encompass broader dimensions of trustworthiness, ethical compliance, and regulatory alignment Evaluation informs decisions regarding model readiness, acceptable residual risk, required mitigation measures, and the appropriate degree of human oversight (ISO, 2018).

Consistent with recent scholarship emphasizing risk based human oversight (Kandikatla & Radeljić, 2025), the paper argues that responsible AI deployment cannot be reduced to performance metrics alone but must be supported by context-sensitive oversight proportionate to the risks revealed through evaluation. Building on the principles of Trustworthy AI (EU, 2019), this paper advances a structured, risk-based approach for integrating Human-in-Command (HIC), Human-in-the-Loop (HITL), and Human-on-the-Loop (HOTL) oversight models into AI governance through the combined evaluation of metrics-derived risk and AI System risk. The proposed Performance-Driven Human Oversight Framework (PD-HOF) for AI Systems follows a stepwise methodology. First, AI System risk is independently identified and categorized based on Model Influence and decision consequence. Second, Metrics are assessed to establish foundational evidence of model performance and statistical validity within the defined context of use, confirming whether predefined functional and acceptance criteria are met. Third, results from metrics are synthesized and categorized into low, medium, or high metrics risk, reflecting residual uncertainty and potential performance-related failure modes. Fourth, metrics risk is mapped against AI system risk using a risk matrix to derive an Integrated AI risk classification. This integrated risk outcome directly informs the proportional allocation of human oversight HIC, HITL, or HOTL ensuring that human involvement is commensurate with both performance uncertainty and potential impact.

Human-in-Command (HIC) provides governance-level accountability, granting authority to suspend or override system operations. Human-in-the-Loop (HITL) enables real-time monitoring, allowing timely human intervention to prevent or mitigate harm. Human-on-the-Loop (HOTL) is suitable when continuous oversight is unnecessary, and monitoring with periodic review is sufficient. By linking performance evidence, contextual risk, and oversight intensity, the proposed PD-HOF framework translates ethical principles into auditable governance mechanisms, while advancing contemporary metrics-informed oversight research (Figure 1).

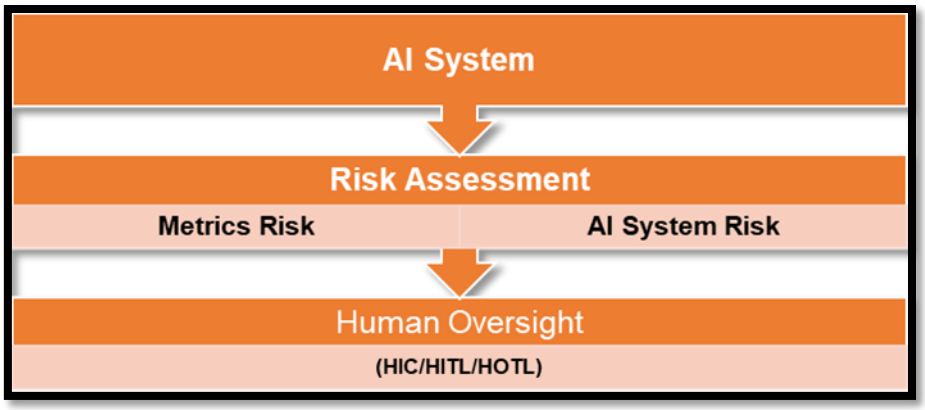

Figure 1. Performance-Driven Human Oversight Framework for AI Systems (Authors' proposal)

Throughout this paper, AI systems encompass large language models (LLMs), Small language models (SLMs), and AI-driven Systems used to support or execute clinical operations within regulated healthcare and research contexts.

2. AI EVALUATION METRICS

2.1. PRIMARY METRICS FOR AI SYSTEMS: QUANTITATIVE ASSESSMENT

Primary metrics constitute the core quantitative measures used to evaluate whether an artificial intelligence (AI) system fulfills its intended functional objectives. These metrics directly assess task-level performance and model correctness and are therefore central to determining baseline system reliability. Among primary metrics, Task Success Rate (TSR) or Task Completion Rate serves as an overarching indicator of system effectiveness, measuring the proportion of end-to-end tasks successfully completed under defined operating conditions. In addition to TSR, commonly used primary metrics include accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC–ROC) for classification tasks, as well as mean squared error (MSE) or related loss functions for regression tasks. (Kocak et al., 2025; McDermott et al., 2024; Musaa & Elnoor, 2025).

Primary performance metrics are quantitatively derived from a confusion matrix, which summarizes model predictions against a clinically established ground truth (Table 1).

Primary Metric	Definition	Interpretation	Formula
Task Completion Rate (TCR)	Proportion of end-to-end clinical tasks successfully executed without system failure or human intervention	Assesses operational reliability and workflow readiness	Tasks completed / Tasks initiated
Accuracy	Proportion of correct predictions among all predictions	Overall correctness; may be misleading in imbalanced datasets	(TP + TN) / (TP + TN + FP + FN)
Precision	Proportion of predicted positives that are truly positive	Controls false positives; relevant for minimizing unnecessary referrals	TP / (TP + FP)
Recall (Sensitivity)	Proportion of actual positive cases correctly identified	Controls false negatives; critical in screening and safety-critical use cases	TP / (TP + FN)
F1 Score	Harmonic mean of precision and recall	Balances false positives and false negatives in imbalanced data	2 × (Precision × Recall) / (Precision + Recall)
AUC–ROC	Ability of the model to discriminate between positive and negative classes across all thresholds	Measures robustness and comparative model performance	Threshold-independent

Table 1. Primary Metrics

TP - True Positive: Cases correctly identified as belonging to the target class

FP - False Positive: Cases incorrectly identified as belonging to the target class

TN - True Negative: Cases correctly identified as not belonging to the target class

FN - False Negative: Cases incorrectly identified as not belonging to the target class

2.2 SECONDARY METRICS FOR AI SYSTEMS: TRUSTWORTHINESS AND QUALITATIVE ASSESSMENT

While primary metrics quantify an AI system's technical and task-level performance, they are insufficient for assessing broader dimensions of trustworthiness, ethical integrity, and real-world reliability. Secondary metrics therefore serve as complementary qualitative and semi-quantitative indicators that evaluate how responsibly and consistently an AI system operates within its intended socio-technical context (Table 2). These metrics are central to governance frameworks such as the EU Guidelines for Trustworthy AI (2019) and ISO/IEC 42001:2023 (AI Management Systems), which emphasize transparency, fairness, robustness, and explainability across the AI lifecycle. By extending evaluation beyond performance alone, secondary metrics enable a more holistic assessment of AI behavior aligned with responsible innovation and regulatory expectations (Gerke et al., 2020; Kim et al., 2023; Kelly et al., 2019; Kowald et al, 2024; Mikalef et al., 2022).

Transparency refers to the degree to which stakeholders can understand an AI system's purpose, design, data usage, and decision-making process. Indicators of transparency include the availability of comprehensive documentation describing data sources, training procedures, underlying assumptions, and model limitations, as well as disclosure of the system's architecture, version history, and changes over time. Clear communication of performance metrics and intended use cases further supports transparency. By promoting accountability and informed trust, transparency enables users, regulators, and auditors to effectively oversee AI throughout its lifecycle. Methodologically, transparency is assessed through structured artifacts such as model cards, datasheets for datasets. (Cheong, 2024; EU, 2019; ISO, 2023a; Mitchell et al., 2019)

Bias and fairness metrics evaluate the consistency and equity of AI predictions across demographic groups, protected attributes, or other data subpopulations. Common indicators include disparities in false positive or false negative rates across groups, representation biases in training or validation datasets, and established fairness measures such as demographic parity, equal opportunity difference, or the disparate impact ratio. These metrics are crucial for identifying and mitigating discriminatory outcomes, ensuring equitable treatment, and reducing ethical, legal, and reputational risks associated with AI deployment. Methodologies for assessing bias and fairness often involve subgroup performance analyses, statistical parity tests, and audits (Bellamy et al., 2019; Castelnovo et al., 2022; Ferrara, 2024).

Robustness captures an AI system's ability to maintain stable and acceptable performance under conditions of data variability, environmental change, or adversarial perturbations. Indicators of robustness include performance degradation when inputs are noisy, incomplete, or modified, sensitivity to outliers or shifts in data distribution, and consistency of performance across temporal, geographical, or contextual variations. Robustness ensures that AI systems perform reliably in dynamic real-world conditions, which may deviate from development environments. Methodologically, robustness is evaluated using stress testing, adversarial assessments, and data drift detection techniques (Mamun et al., 2025)

Explainability refers to the extent to which AI outputs can be meaningfully interpreted and justified by human stakeholders. Explainability supports user understanding, decision traceability, and accountability, particularly in high-impact or regulated applications. Assessment methodologies typically combine model-agnostic explanation techniques with human-centered evaluation, including expert panels or structured user feedback, to ensure clarity, relevance, and interpretability (Baker & Xiang, 2023; EU, 2019).

Additional secondary metrics are necessary. Hallucination evaluates the tendency of a system to generate factually incorrect, misleading, or unsupported outputs, which can amplify downstream clinical, operational, or regulatory risk (Aljamaan et al., 2024). Consistency, treated as a cross-cutting attribute, measures output stability across repeated executions of the same task and is especially critical in GxP and other safety-critical environments where non-deterministic behavior may undermine process control. From a risk management perspective, deficiencies in secondary metrics increase uncertainty, reduce detectability of failure modes, and elevate overall metrics risk, consistent with the principles of ICH Q9 and ISO/IEC 23894 (ICH, 2023; ISO, 2023; EU, 2019).

Collectively, these secondary metrics provide structured, complementary insights that extend AI evaluation beyond technical performance to encompass trustworthiness, ethical operation, and operational reliability. They form an essential component of a risk-based framework for AI validation, ensuring that systems meet both functional and governance expectations in regulated or high-stakes environments.

Secondary Metric	Description
Transparency	Availability and completeness of model and data documentation. Disclosure of intended use, limitations, and versioning.
Bias & Fairness	Fairness scores (demographic parity, equal opportunity difference, disparate impact ratio)
Robustness	Stress Testing, adversarial tests and data drift.
Explainability	Local explanations showing feature contributions to individual predictions. Global feature importance distributions Stability of explanations across samples
Consistency	Variability in outputs across repeated runs or similar inputs Non-deterministic behavior patterns
Hallucination	Frequency of unsupported or incorrect outputs

3. RISK ASSESSMENT AND OVERSIGHT

Risk assessment is a structured approach to identifying, analyzing, and evaluating risks that may impact system quality, patient safety, data integrity, or regulatory compliance. It is required to support risk-based decision-making by categorizing risks and enabling the selection and implementation of appropriate controls that are commensurate with the level of risk (ICH, 2023; ISO, 2018; ISO/IEC, 2023). This approach introduces a metrics-based risk assessment in which primary and secondary performance metrics are systematically evaluated to establish an initial risk category (low, medium, or high). The resulting metrics risk is subsequently integrated with the inherent risk of the AI system to determine an overall risk classification. This combined risk assessment (Figure 2) informs the selection and application of proportionate human oversight controls.

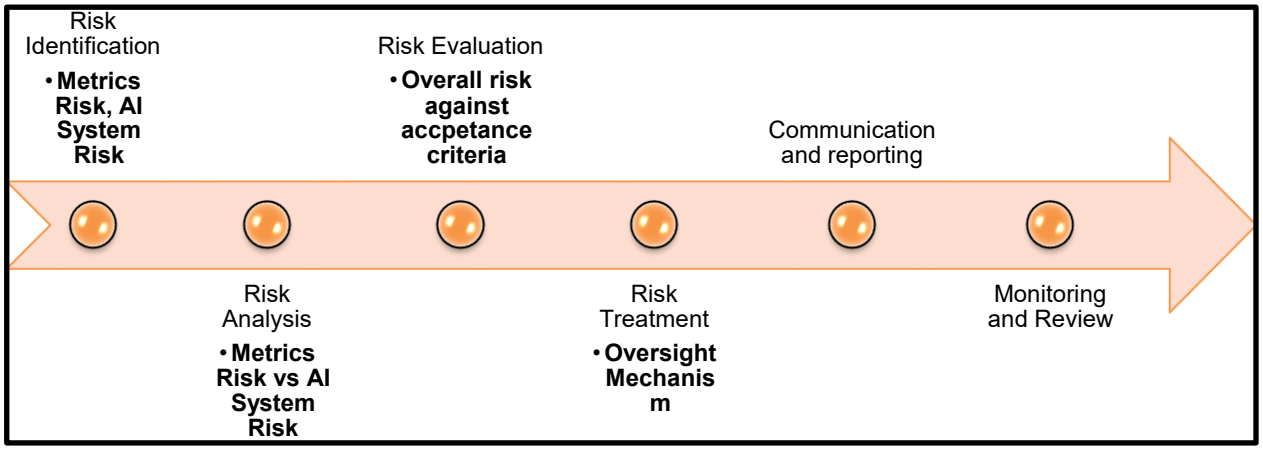

Figure 2. Risk assessment process mapping with ISO 31000 (Authors proposal)

3.1 RISK IDENTIFICATION

AI System risk

AI System risk represents the inherent risk associated with an AI system based on its intended context of use, its role in decision-making, and its impact on regulated GxP activities. Consistent with ICH Q9 Quality Risk Management, AI System risk focuses on potential effects on patient safety, product quality, and data integrity, rather than on the underlying technology itself. Evaluating AI System risk is critical for ensuring that AI systems deployed in regulated or clinical environments operate safely, reliably, and in compliance with applicable standards (ASME, 2018; FDA, 2025; ICH, 2023; ISO, 2018; ISO/IEC, 2023).

AI system risk is typically assessed across three interrelated dimensions: clinical context, decision consequence, and GxP impact. The clinical context reflects the therapeutic area and intended purpose of the AI system, such as diagnostic support, predictive analytics, trial optimization, or administrative automation (ASME, 2018; FDA, 2025; ICH, 2023; ISO, 2018; ISO/IEC, 2023). This dimension establishes a baseline for criticality, as different use cases inherently carry varying risk profiles. For example, AI systems that support clinical decision-making or patient stratification present higher inherent risk than those used for operational or administrative tasks. Decision consequence evaluates the potential severity of harm if AI outputs are incorrect, misleading, delayed, or unavailable. This dimension aligns with the severity aspect of ICH Q9 and considers impacts such as inappropriate clinical decisions, protocol deviations, delayed interventions, or regulatory non-compliance. GxP impact assesses the extent to which the AI system influences regulated activities, including patient safety, data integrity, product quality, and regulatory compliance. Systems that directly affect clinical outcomes, generate or transform GxP data, or support critical quality decisions are assigned to higher GxP impact (ASME, 2018; FDA, 2025; ICH, 2023; ISO, 2018; ISO/IEC, 2023).

Primary Metrics Risk

Primary metrics such as accuracy, precision, recall (sensitivity), F1 score, AUC–ROC, and task completion rate (TCR) quantify the technical performance of an AI system. Authors' proposal is to map these metrics into a risk-based framework, where lower performance indicates a higher probability of failure, increased uncertainty, or potential adverse outcomes. This approach allows organizations to prioritize mitigation, validation, and monitoring efforts based on risk-informed decisions. Each primary metric will be mapped to Low, Medium, High risk levels using performance thresholds (Table 3).

Low Risk: AI performance meets or exceeds the expected performance thresholds. System behavior is highly reliable, minimal probability of error or adverse outcome.

Medium Risk: AI performance is adequate but not optimal. Moderate probability of errors or unintended outcomes; additional validation, monitoring, or mitigation may be required.

High Risk: AI performance is suboptimal; high probability of errors or failure; immediate attention, re-training, or operational restrictions may be needed.

Primary Metric	High Performance = Low Risk	Medium Performance = Medium Risk	Low Performance = High Risk
Task Completion Rate (TCR)	Greater than 95%	85 to 94%	80 - 85%
Accuracy	Greater than 90 %	80 to 89 %	75 % to 80 %
Precision	Greater than 90 %	80 to 89%	75 % to 80 %
Recall (Sensitivity)	Greater than 90 %	80 to 89%	75 % to 80 %
F1 Score	Greater than 0.85	0.75 to 0.84	0.70 to 0.75
AUC–ROC	Greater than 0.90	0.80 to 0.89	0.70 to 0.75

Table 3. Primary Metrics Risk Classification (Author's proposal)

The exact thresholds should be tailored to the clinical or operational domain, considering the consequences of errors (e.g., missed diagnoses in healthcare are higher risk than false positives).

Secondary Metrics Risk

Secondary Metrics such as Transparency, Fairness, Robustness, Explainability serve as complementary qualitative and semi-quantitative indicators that evaluate how responsibly and consistently an AI system operates. Authors proposal is to map each secondary metric to Low, Medium, High risk levels using indicators such as documentation completeness, parity in error rates, model performance stability, interpretability score (Table 4). Documenting rationale and measurement method for each rating. This ensures reproducibility and auditability.

Low Risk: Meets or exceeds all defined criteria and indicates strong performance, trustworthiness, or compliance. Medium Risk: Partially meets criteria, moderate risk or uncertainty. High Risk: Does not meet criteria, high risk, gaps in trustworthiness, or non-compliance

Metric	High = Low risk	Medium = Medium Risk	Low = High Risk
Transparency	Comprehensive documentation, model card & datasheets available, clear description of architecture, data, and limitations	Partial documentation, some missing details, unclear communication of use or limitations	Minimal/no documentation, architecture or data sources unclear; stakeholders cannot interpret AI decisions

Bias and Fairness	Minimal and clinically non-meaningful disparities across groups; fairness metrics within accepted thresholds; no systematic bias detected.	Moderate disparities observed in certain subgroups but not critical	Significant disparities across groups; high risk of discrimination or inequitable outcomes
Robustness	Stable performance across noisy inputs, data drift, or adversarial tests.	Moderate degradation under stress conditions; partially mitigated	High sensitivity to noise, outliers, or environmental changes; performance unreliable
Explainability	Explanations are stable, interpretable, and aligned with domain knowledge; understandable to intended users	Explanations partially interpretable; alignment with domain knowledge is inconsistent	Explanations unavailable, incomprehensible, or misaligned with domain knowledge
Consistency	Outputs are stable and repeatable across runs and similar inputs.	Moderate variability observed, requiring monitoring or controls	High instability or unpredictability, especially in critical decisions
Hallucination	Outputs are consistently factual and verifiable	Occasional unsupported content with detectable errors	Frequent hallucinations or unverifiable

Table 4. Secondary Metrics Risk Classification (Author's proposal)

Thresholds should be context-specific, depending on clinical relevance, regulatory requirements, and task criticality.

3.2. RISK ANALYSIS

AI system Risk

To operationalize AI System risk, a risk matrix is established, with Model influence against decision consequence (ASME, 2018; US FDA, 2025). Model influence reflects the system's influence on regulated activities, categorized as Low, Medium, or High, depending on its criticality in processes affecting patient safety, data integrity, product quality, or regulatory compliance (ASME, 2018; FDA, 2025; ICH, 2025). Decision consequence represents the combined assessment of outcome severity, likelihood of occurrence, and detectability of harm arising from incorrect, incomplete, delayed, or unavailable AI outputs. Based on this integrated evaluation, decision consequence is categorized as low, medium, or high (ASME, 2018; FDA, 2025; ICH, 2025). The intersection of these two dimensions yields an AI system risk category (Figure 3).

Low Risk: Systems with limited impact on regulated activities and minimal potential harm if outputs are incorrect or delayed. Medium Risk: Systems with moderate influence on GxP processes where errors could result in manageable but non-trivial consequences. High Risk: Systems with substantial impact on patient safety, data integrity, or product quality, where incorrect or delayed outputs could cause serious harm or regulatory non-compliance.

	Decision Consequence			
	Risk Matrix	**Low**	**Medium**	**High**
Model Influence	**High**	Medium	High	High
	Medium	Low	Medium	High
	Low	Low	Low	Medium

Figure 3. AI System Risk assessment matrix (Author's proposal)

This matrix-based approach ensures that AI system risk is evaluated in a consistent, transparent, and reproducible manner, aligning with ICH Q9's principles of proportionality and science-based risk assessment.

Integrated Risk Matrix Methodology: Metrics Risk and AI System Risk

To enable a comprehensive and risk-based evaluation of AI systems, overall risk can be determined by combining Metrics Risk derived from metrics with AI System Risk, which reflects the system's role in regulated GxP activities, clinical context, and decision consequences. This approach recognizes that AI system risk is influenced both by the system's technical reliability and its potential impact on patient safety, product quality, and regulatory compliance.

	AI System Risk			
	Risk Matrix	Low	Medium	High
Metrics Risk	**High**	Medium	High	Critical
	Medium	Low	Medium	High
	Low	Low	Low	Medium

Figure 4. Integrated Risk assessment matrix (Author's proposal)

The intersection of these dimensions (Metrics Risk and AI system risk) produces integrated risk matrix, as illustrated in the matrix (Figure 5).

Systems exhibiting high risk across both metrics risk and AI system risk warrant a halt in progression and require further investigation, remediation, and revalidation prior to continuation.

Given the presence of multiple evaluation metrics, metrics risk is determined using a conservative approach, either by selecting the worst-case outcome across all assessed metrics, by prioritizing predefined significant metrics, or as otherwise documented.

Supplementary Approach:
Metrics Risk can also be systematically assessed by combining primary and secondary metrics within a risk matrix framework. In this approach, the risk associated with the risk determined from primary metrics, including accuracy, precision, recall, F1 score, AUC–ROC, and task completion rate is evaluated in conjunction with the secondary metrics such as transparency, fairness, robustness, explainability, consistency, and hallucination propensity. The resulting matrix allows each combination of primary and secondary metrics performance to be mapped to an overall risk category (Figure 4).

	Secondary Metrics Risk			
	Risk Matrix	Low	Medium	High
Primary Metrics Risk	**High**	Medium	High	Critical
	Medium	Low	Medium	High
	Low	Low	Low	Medium

Figure 5. Metrics Risk assessment matrix (Author's proposal)

The interaction between primary and secondary metrics risk is evaluated using a simple risk matrix to determine the overall metrics risk. Systems exhibiting high risk across both primary and secondary metrics warrant a halt in progression and require further investigation, remediation, and revalidation prior to continuation.

3.3. RISK EVALUATION

Risk evaluation involves comparing analyzed risks against predefined acceptance criteria to prioritize AI systems into low, medium, or high-risk categories (ICH, 2023; ISO, 2018). This prioritization informs decisions regarding deployment readiness, validation depth, and the necessity for additional controls. Where risks need to be reduced to acceptable thresholds for safer deployment and operation, risk treatment and control measures are selected and implemented with a particular focus on proportionate human oversight mechanisms, including Human-in-Command (HIC), Human-in-the-Loop (HITL), or Human-on-the-Loop (HOTL), selected according to the assessed risk level (Kandikatla & Radeljić, 2025).
Continuation of body – after source code.

3.4. RISK TREATMENT

Once the integrated risk is determined by integrating Metrics Risk and AI system Risk, the level of human oversight can be proportionally assigned according to the system's risk priority (ISO, 2018; ISO, 2023) (Figure 6). High-risk AI systems, where critical metrics are deficient or the AI System has substantial impact on patient safety, product quality, or regulatory compliance, require a Human-in-Command (HIC) model. In this model, a qualified human retains ultimate authority and responsibility over decisions, ensuring that AI outputs are carefully interpreted before action. Medium-risk systems are best managed under a Human-in-the-Loop (HITL) framework, in which human operators review and validate AI outputs during operation, allowing for corrective intervention when necessary. Low-risk systems, with minimal potential impact or robust technical performance, can be monitored under a Human-on-the-Loop (HOTL) model, where AI operates autonomously while human supervision occurs primarily through auditing, alerts, or periodic review (Table 5). By linking risk classification to oversight intensity, this framework ensures that human intervention is applied proportionally to both technical and operational risk, supporting safe, compliant, and responsible deployment of AI in clinical and regulated environments (Kandikatla & Radeljić, 2025).

Risk Priority	Oversight Model	Description
High	Human-in-Command (HIC)	Humans retain ultimate authority and responsibility over decisions. Used in critical systems affecting diagnosis or treatment decisions.
Medium	Human-in-the-Loop (HITL)	Humans continuously monitor and can intervene, or override AI outputs as needed. Common in supportive decision systems.
Low	Human-on-the-Loop (HOTL)	Humans establish operational parameters and conduct periodic audits, but continuous intervention is not required. Suitable for low-impact or mature AI systems.

Table 5. Risk Priority and Human Oversight

3.5 COMMUNICATION & REPORTING

Communication and reporting constitute a critical cross-cutting activity throughout the risk management process. Risk assumptions, assessment methodologies, metric results, risk categorizations, and implemented control measures are formally documented and communicated to relevant stakeholders, including developers, quality and compliance teams, clinical users, and regulatory authorities. This documentation supports traceability, auditability, and regulatory transparency, enabling informed decision-making and accountability across the AI lifecycle. (ICH, 2023; ISO, 2018; ISO/IEC, 2023).

3.6 MONITORING & REVIEW

Finally, monitoring and review ensures that AI risk management remains effective over time. Post-deployment monitoring activities track residual risks, assess the ongoing effectiveness of mitigation controls, and detect emerging risks arising from model updates, data drift, changes in clinical practice, or evolving regulatory expectations. Periodic review of metrics enables timely reclassification of risk levels and adjustment of oversight mechanisms, thereby supporting continuous improvement and sustained compliance with ISO and ICH quality risk management principles (ICH, 2023; ISO, 2018; ISO/IEC, 2023).

4. EXAMPLE

Clinical data quality management agent

The AI system risk is evaluated as medium, taking into account both Model Influence and decision consequences. The metrics risk is likewise assessed as Low, based on the Metrics Risk. By mapping AI system risk (High) against metrics risk (Low) the resulting overall risk level is medium, consistent with the 3×3 risk matrix approach.

Type	Result
AI System Risk	**Medium**
Model Influence	*High*
Decision Consequence	*Low*
Metrics Risk	**Low**
Metrics risk	*Low (e.g. TCR = 96 %)*
Overall Risk	**Medium**
Human Oversight	**HITL**

5. CONCLUSION

AI evaluation and risk assessment are essential for ensuring safe, reliable, and ethically responsible deployment, particularly in high-impact domains such as healthcare. By systematically assessing primary and secondary metrics, organizations can quantify metrics risk and map it against AI System risk, defined by GxP impact and decision consequence. This combined assessment informs the overall risk level, guiding the appropriate human oversight model—Human-in-the-Loop (HITL), Human-on-the-Loop (HOTL), or Human-in-Command (HIC).

Risk evaluation involves identifying both metrics and system risks, while risk treatment implements controls and oversight strategies tailored to the risk level. Monitoring and review ensure that residual risks are managed, system performance remains within acceptable bounds, and AI outputs continue to meet clinical or operational standards. Communication and reporting provide transparency to stakeholders, regulators, and auditors, reinforcing accountability and trust.

The framework (PD-HOF) is designed to be flexible and adaptable and supports multiple evaluation pathways to inform human oversight decisions. (i) Evaluating agent risk independently to understand baseline oversight requirements. (ii) Comparing AI system risk with primary metrics risk to assess how core performance reliability influences oversight needs. (iii) Comparing AI system risk with secondary metrics risk to examine the impact of supporting or contextual metrics on oversight. (iii) Evaluating primary versus secondary metrics risk to determine how differing metric sensitivities affect oversight intensity.

It prioritizes the most significant or worst-case metrics when evaluating risk, accommodates variations based on the AI system or context of use, and allows human oversight to be applied based on either metrics risk or system risk independently. Risk score ranges can also be adjusted according to organizational goals and safety priorities. Overall, this approach provides a robust, risk-informed methodology for AI validation, deployment, and governance, balancing technical performance with ethical, regulatory, and operational considerations

REFERENCES

Aljamaan, F., Temsah, M. H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., ... & Malki, K. H. (2024). Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, *12*(1), e54345.

ASME. (2018) V&V440-Assessing credibility of computational modeling through verification and validation: Application to medical devices. https://www.asme.org/codes-standards/find-codes-standards/assessing-credibility-of-computational-modeling-through-verification-and-validation-application-to-medical-devices

Baker, S., & Xiang, W. (2023). Explainable AI is responsible AI: How explainability creates trustworthy and socially responsible artificial intelligence. *arXiv preprint arXiv:2312.01555*.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4-1.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape Sci.

Cheong, B. C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1421273.

European Commission, S. (2019). Ethics guidelines for trustworthy AI. *Publications Office*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295-336). Academic Press.

Google Developers. (2025). Classification: Accuracy, recall, precision, and related metrics. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall
Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681.
International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2023). ICH Q9(R1): Quality risk management. https://www.ich.org
International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2025). ICH E6(R3): Guideline for good clinical practice. https://database.ich.org
ISO. (2023). ISO/IEC 23894: https://www.iso.org/standard/77304.html
ISO. (2023a). ISO/IEC 42001. https://www.iso.org/standard/42001
ISO. (2018). ISO 31000. https://www.iso.org/standard/65694.html#lifecycle
Kandikatla, L., & Radeljic, B. (2025). AI and Human Oversight: A Risk-Based Framework for Alignment. *arXiv preprint arXiv:2510.09090*.
Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195. https://doi.org/10.1186/s12916-019-1426-2
Kim, M., Sohn, H., Choi, S., & Kim, S. (2023). Requirements for trustworthy artificial intelligence and its application in healthcare. *Healthcare Informatics Research*, *29*(4), 315-322.
Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, *7*, 1456486. https://doi.org/10.3389/frai.2024.1456486 Frontiers
Kocak, B., Klontzas, M. E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., ... & Cuocolo, R. (2025). Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, 100030.
Kowald, D., Scher, S., Pammer-Schindler, V., Müllner, P., Waxnegger, K., Demelius, L., ... & Kopeinik, S. (2024). Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data*, *7*, 1467222.
Mamun, A., Soumma, S. B., & Ghasemzadeh, H. (2025). Trustworthy AI in Digital Health: A Comprehensive Review of Robustness and Explainability.
McDermott, M., Zhang, H., Hansen, L., Angelotti, G., & Gallifant, J. (2024). A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, *37*, 44102-44163.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). https://doi.org/10.1145/3287560.3287596
Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side'of AI. European Journal of Information Systems, 31(3), 257-268.
Nguyen, L. P. T., & Do, H. T. (2025, December). RAISE: A Unified Framework for Responsible AI Scoring and Evaluation. In *International Conference on Principles and Practice of Multi-Agent Systems* (pp. 453-460). Cham: Springer Nature Switzerland.
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
Ripatti, A. R. (2025). Screening Diabetic Retinopathy and Age-Related Macular Degeneration with Artificial Intelligence: New Innovations for Eye Care.
US FDA. (2025). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. https://www.fda.gov/media/168399/download
Uy, H., Fielding, C., Hohlfeld, A., Ochodo, E., Opare, A., Mukonda, E., ... & Engel, M. E. (2023). Diagnostic test accuracy of artificial intelligence in screening for referable diabetic retinopathy in real-world settings: A systematic review and meta-analysis. *PLOS Global Public Health*, *3*(9), e0002160.

CONTACT INFORMATION

Author Name: Laxmiraju Kandikatla
Company: Maxis AI
Address: 510 Thornall street 180, Edison, NJ 08837
Work Phone: NA
Email: laxmiraju.k@maxisit.com
Website: www.maxisit.com