Trust Architectures for AI-Enabled Drug Discovery: Transparency, Validation, Reproducibility, Accountability, and Regulatory Alignment

Saif Al-Hinai[1]*, Amal Al-Balushi[1], Mohammed Al-Maskari[2]

[1]Department of AI for Frankincense and Myrrh Therapeutic Potential, College of Pharmacy, Sultan Qaboos University, Muscat, Oman.
[2]Department of Computational ADME for Volatile Oils, College of Pharmacy, University of Nizwa, Nizwa, Oman.

ABSTRACT

Artificial intelligence is increasingly embedded in drug discovery through molecular representation, target identification, property prediction, virtual screening, candidate prioritization, and evidence interpretation. However, predictive performance alone cannot establish that an AI system is trustworthy, scientifically credible, or suitable for consequential decision support. This article proposes a conceptual trust architecture that organizes technical, evidentiary, organizational, and lifecycle requirements for AI-enabled drug discovery. The architecture treats transparency as a structured combination of data provenance, dataset documentation, model documentation, traceability, interpretability, explainability, and uncertainty reporting. Validation is organized through predefined claims, context-sensitive benchmarks, leakage controls, internal and external evaluation, and prospective or experimental assessment where relevant. Reproducibility is supported through documented workflows, version control, accessible computational resources where appropriate, and reconstructable model-development records. Accountability is addressed through human oversight, assigned governance roles, auditable decisions, bias and fairness assessment, risk management, and explicit escalation boundaries. Regulatory alignment is framed as the organization of evidence, monitoring, change control, and implementation safeguards rather than evidence of regulatory approval or legal sufficiency. The proposed architecture also incorporates lifecycle monitoring, drift detection, controlled model updating, and feedback governance. Its principal contribution is an integrated framework for connecting trustworthy-AI principles with the practical evidence and decision structures of drug discovery while maintaining clear boundaries between governance readiness, implementation readiness, clinical utility, and regulatory authorization.

Key Words: *Trustworthy AI, Drug discovery, Governance architecture, Transparency, Validation, Reproducibility*

eIJPPR 2026; 16(2):106-122

HOW TO CITE THIS ARTICLE: Al-Hinai S, Al-Balushi A, Al-Maskari M. Trust Architectures for AI-Enabled Drug Discovery: Transparency, Validation, Reproducibility, Accountability, and Regulatory Alignment. Int J Pharm Phytopharmacol Res. 2026;16(2):106-22. https://doi.org/10.51847/nI2B8SaJLi

INTRODUCTION

Artificial intelligence has expanded across drug discovery and development, supporting activities such as target identification, molecular-property prediction, structure-based and ligand-based screening, de novo design, biomarker analysis, candidate prioritization, and clinical-development planning. These applications depend on heterogeneous molecular, biological, experimental, and clinical data, together with computational pipelines that transform those inputs into rankings, predictions, classifications, or decision-support outputs. The breadth of these applications makes AI relevant across the pharmaceutical research lifecycle, but it also means that the credibility of an output depends on more than the apparent performance of an isolated model [1].

Expectations surrounding AI-enabled drug discovery have often exceeded the strength of the available evidence. A

Corresponding author: Saif Al-Hinai
Address: Department of AI for Frankincense and Myrrh Therapeutic Potential, College of Pharmacy, Sultan Qaboos University, Muscat, Oman.
E-mail: ✉ saif.alhinai@squ.edu.om
Received: 05 February 2026; **Revised:** 14 April 2026; **Accepted:** 15 April 2026

model may achieve favorable results within a selected benchmark while remaining vulnerable to dataset artifacts, unrealistic evaluation conditions, narrow applicability domains, or weak correspondence with prospective discovery problems. Distinguishing plausible impact from technological overstatement therefore requires careful attention to the claims being made, the datasets and comparators used, and the degree to which evaluation conditions resemble the intended scientific context [2].

The quality and meaning of chemical and biological data create additional trust challenges. Molecular datasets may combine results from different assay formats, laboratories, measurement conditions, annotation procedures, and biological contexts. Labels that appear computationally interchangeable may represent distinct experimental phenomena, while missing data, selective reporting, duplicated chemical series, and unrecorded preprocessing can distort model learning. These limitations constrain validity, generalizability, and biological interpretation unless provenance, representativeness, and intended-use boundaries are made explicit [3].

Governance is consequently required across the full range of AI applications rather than being confined to model selection or final performance reporting. Drug discovery teams need a structure that connects data stewardship, model documentation, validation planning, reproducibility, accountability, risk management, lifecycle monitoring, and human review. This article therefore proposes a trust architecture for AI-enabled drug discovery that organizes these functions into linked evidence and oversight layers. The architecture is intended to support scientific credibility and governance readiness while avoiding claims that documentation, explainability, validation, or regulatory alignment independently establish deployment readiness, clinical utility, legal sufficiency, or regulatory approval [4].

Trust challenges in AI-enabled drug discovery

Opacity is a central challenge when complex models influence compound selection, target prioritization, or experimental-resource allocation. Post-hoc explanations may help users inspect associations or identify influential features, but they do not necessarily reveal the model's actual decision process or demonstrate that the learned relationship is scientifically valid. In consequential settings, interpretability should therefore be considered a design and governance question rather than an optional visualization step, particularly where a simpler and more understandable model could perform the required function [5].

Trustworthiness is also limited by gaps between model development and the context in which an output is expected to influence research. Internal performance may not persist across institutions, laboratories, assay platforms, chemical series, disease subtypes, or future data. Workflow integration, external evaluation, stakeholder understanding, data quality, and operational oversight can therefore be as important as algorithmic accuracy. A technically successful model may still be unsuitable for decision support when its intended users, failure modes, monitoring requirements, or escalation procedures have not been defined [6].

Benchmark design can produce a misleading impression of generalization. In ligand-based modeling, random data partitioning may distribute structurally similar compounds across training and test sets, allowing models to exploit recurring chemical patterns rather than infer relationships that transfer to genuinely novel molecular series. Under these conditions, apparently strong performance may reflect memorization or interpolation within familiar chemical neighborhoods. Benchmark credibility consequently depends on split design, similarity analysis, comparator selection, applicability-domain assessment, and alignment between the benchmark task and the intended discovery claim [7].

Bias can also enter through the construction of chemical and structural datasets. Artificial distinctions between active and inactive compounds, systematic differences in molecular properties, duplicated scaffolds, selection effects, or benchmark-specific artifacts may allow a model to distinguish dataset origin rather than biologically meaningful activity. Such biases weaken generalizability and can remain concealed when only aggregate performance metrics are reported. Trust challenges therefore extend beyond opacity to include incomplete provenance, leakage, uncertainty, limited representativeness, ambiguous responsibility, lifecycle change, and unclear regulatory relevance. These interacting risks explain why model performance alone cannot constitute evidence of trustworthiness [8].

Transparency and validation

Transparency begins with reconstructable evidence lineage. Dataset records should describe where information originated, how compounds or biological observations were selected, which transformations were applied, what exclusions occurred, and which intended or prohibited uses apply. Structured dataset documentation can make composition, collection processes, known limitations, maintenance practices, and potential sources of bias visible to reviewers and downstream users [9]. Model documentation should complement this record by specifying the intended function, architecture, inputs, outputs, training procedure, optimization choices, software environment, evaluation design, limitations, responsible owners, and version history. Reporting frameworks can

strengthen this process by defining minimum information needed to understand how a biomedical AI model was developed and assessed [10].

Explainability and interpretability should be separated conceptually and operationally. Interpretability concerns whether relevant users can understand the model's decision logic, whereas explainability may involve additional methods that summarize, approximate, or visualize the basis of a prediction. The appropriate form of explanation depends on the stakeholder, scientific purpose, decision consequence, and type of uncertainty involved [11]. Explanations should therefore be evaluated for fidelity, stability, relevance, and susceptibility to misleading interpretation. They may support investigation and communication, but they do not independently demonstrate predictive validity, biological correctness, fairness, safety, or trustworthy use [12].

Validation should begin with a predefined plan that links every evaluation to a specific claim, intended use, context, endpoint, comparator, acceptance criterion, and failure condition. In biological machine learning, validation quality depends on transparent treatment of the data, optimization process, model specification, and evaluation strategy [13]. Internal validation should test overfitting, tuning bias, calibration, and sensitivity within appropriately separated development data. External validation should examine transportability to independent chemical series, laboratories, assays, populations, institutions, or time periods. Prospective validation, where relevant, should evaluate future cases under a protocol defined before outcomes are observed, while impact-oriented evaluation should remain distinct from technical performance assessment [14].

Uncertainty reporting is necessary because predictions differ in calibration, applicability, and evidentiary strength. Probabilistic outputs should be evaluated for agreement between predicted and observed outcomes, and material miscalibration should trigger reassessment or updating rather than being obscured by discrimination metrics alone [15]. Benchmark evaluation should also disclose chemical similarity, split rationale, baseline selection, contamination risks, and the extent to which results may reflect interpolation rather than generalization [7]. Data-leakage controls should cover preprocessing, feature construction, entity duplication, temporal separation, hyperparameter selection, and all pathways through which evaluation information could influence model development [8]. Where computational predictions are intended to support compound or mechanism selection, the validation plan should additionally specify when orthogonal experimental confirmation is required, while recognizing that confirmation of a limited endpoint does not establish safety, therapeutic efficacy, or clinical utility [1]. **Table 1** summarises transparency and validation requirements for trustworthy AI-enabled drug discovery.

Table 1. Transparency and Validation Requirements for AI-Enabled Drug Discovery: Data Provenance, Dataset Documentation, Model Documentation, Explainability, Interpretability, Traceability, Validation Planning, Benchmark Quality, Data Leakage Prevention, Uncertainty Reporting, and Evidence Boundaries

Trust component	Core question	Required evidence or practice	Governance function	Failure risk	Validation need	Architecture role	Decision boundary
Data provenance	Where did each data element originate, and how was it modified?	Source identifiers, acquisition context, assay conditions, timestamps, licenses, exclusions, curation decisions, and transformation history.	Establishes an inspectable evidence lineage.	Unrecognized contamination, incompatible measurements, unverifiable labels, or inappropriate reuse.	Provenance records should be versioned, complete, and reconstructable.	Foundational data-governance control.	Data with unresolved material provenance gaps should not support high-consequence claims.
Dataset documentation	What does the dataset represent, and for which uses is it suitable?	Documentation of composition, collection, curation, missingness, class balance, chemical or	Makes suitability and representativeness reviewable.	Hidden imbalance, unsupported extrapolation, and misleading reuse outside the	Documentation should accompany each dataset version and material update.	Dataset evidence artifact.	Documentation supports review but does not itself establish representativeness or validity.

		biological coverage, intended use, limitations, and prohibited use.		sampled domain.			
Model documentation	What was developed, for what purpose, and under which assumptions ?	Intended use, architecture, inputs, outputs, training process, hyperparameters, dependencies, limitations, owners, and version identifiers.	Enables scrutiny, ownership assignment, and change control.	Irreproducible behavior, unclear responsibility, and undocumented model changes.	The documented model must correspond exactly to the evaluated version.	Model evidence artifact.	A documented model is not necessarily a validated or useful model.
Explainability	Can an output be communicated meaningfully to relevant stakeholders ?	Defined explanation target, method rationale, fidelity assessment, stability assessment, limitations, and stakeholder-specific presentation.	Supports review, challenge, investigation, and communication.	Plausible but misleading explanations and false confidence in model reasoning.	Explanation behavior should be assessed on relevant examples and failure cases.	Stakeholder-facing transparency mechanism .	An explanation does not establish causal correctness, validity, safety, or utility.
Interpretability	Is the decision logic understandable without relying entirely on post-hoc approximation?	Interpretable model or representation where feasible, with justification when greater complexity is retained.	Improves direct scrutiny of consequential model behavior.	Hidden shortcuts, unexplained complexity, and inability to challenge decision logic.	Interpretability should be assessed relative to the task, user, and consequence .	Model-design governance control.	Interpretability alone does not establish performance or biological validity.
Traceability	Can each output be linked to the exact data, code, model, settings, and actions that produced it?	Run identifier, data version, code commit, model hash, software environment, parameters, timestamp, user action, and decision record.	Enables reconstruction, audit, and incident investigation.	Inability to reproduce an output or identify the source of an error.	Traceability should be tested through end-to-end reconstruction exercises.	Cross-layer linking mechanism .	Untraceable outputs should not support consequential decisions.
Validation planning	Which claims will be	Predefined intended use, endpoints,	Prevents selective reporting and	Moving thresholds, cherry-	The plan should be established	Validation-gate blueprint.	Validation supports only the claims and

	evaluated, in which contexts, and against what criteria?	datasets, comparators, metrics, uncertainty measures, acceptance criteria, and failure thresholds.	post-hoc expansion of claims.	picked results, and evaluation unrelated to intended use.	before final model assessment.		contexts specified in the plan.
Internal validation	Does performance persist within appropriately separated development data?	Nested resampling, scaffold-aware or temporal partitioning, calibration assessment, sensitivity analysis, tuning isolation, and leakage checks.	Identifies overfitting and unstable estimates.	Optimistic results caused by tuning, correlated samples, or repeated chemical series.	Data partitions and optimization procedures must be reproducible.	Initial technical evidence gate.	Internal validation does not establish external generalizability.
External validation	Does the model transfer to genuinely independent data or settings?	Independent assays, laboratories, chemical series, populations, sites, or time periods, with subgroup and calibration analyses.	Tests transportability beyond development conditions.	Context-specific performance presented as general capability.	Independence and comparability of external evidence should be documented.	Context-transfer evidence gate.	Success in one external setting does not demonstrate universal validity.
Prospective validation where relevant	Does the model retain performance on future observations or decisions?	Prospectively specified protocol, temporal separation, locked evaluation, predefined endpoints, and analysis completed without outcome-informed revision.	Tests forward-looking reliability.	Retrospective artifacts being mistaken for prospective usefulness.	Protocol and analysis should be fixed before relevant outcomes become available.	Translational evidence gate.	Prospective validity does not independently establish clinical utility or approval.
Experimental validation where relevant	Are computational predictions supported by suitable laboratory evidence?	Orthogonal assays, appropriate controls, replication, dose-response assessment, target-engagement	Connects computational predictions with biological evidence.	Candidate prioritization based solely on unconfirmed computational output.	Experimental procedures and results should be reproducible and independently examined	Biological-evidence gate.	Confirmation of a limited endpoint does not establish safety, efficacy, or therapeutic benefit.

		evidence, and predefined experimental criteria.			where feasible.		
Benchmark quality	Does the benchmark assess the intended capability rather than dataset familiarity?	Split rationale, novelty analysis, comparator justification, metric selection, baseline models, applicability-domain assessment, and contamination checks.	Protects the credibility of comparative evaluation.	Memorization, inflated rankings, and misleading claims of discovery capability.	Benchmark construction and evaluation procedures should be reproducible.	Comparative-evidence control.	Superior benchmark performance does not equal prospective discovery success.
Data leakage prevention	Has information improperly crossed between training, tuning, preprocessing, and evaluation?	Entity-level deduplication, preprocessing isolation, temporal controls, scaffold separation, pipeline review, and independent leakage audit.	Protects the integrity of performance estimates.	Severe optimism, false generalization, and invalid comparison.	Leakage analysis must cover the complete computational pipeline.	Mandatory validation checkpoint.	Unresolved leakage invalidates the affected performance claim.
Uncertainty reporting	How reliable is each prediction, and when is the model outside its supported domain?	Calibration measures, confidence or prediction intervals, applicability-domain indicators, error analysis, abstention rules, and uncertainty sources.	Enables risk-sensitive interpretation and escalation.	False precision, inappropriate downstream action, and concealed out-of-domain use.	Uncertainty estimates should be validated on contextually relevant data.	Decision-support constraint.	High uncertainty or unsupported-domain status should trigger abstention or expert review.
Evidence boundaries	What may legitimately be concluded from the available evidence?	Explicit claim limits separating technical performance, reproducibility, experimental support, implementation readiness, clinical utility, and	Prevents unsupported escalation of evidentiary claims.	Governance documentation being misrepresented as proof of safety, effectiveness, compliance, or approval.	Boundaries should be reviewed whenever intended use, data, model, or evidence changes.	Cross-cutting claim-control layer.	Transparency and validation may support trustworthiness but do not establish legal sufficiency, regulatory approval, or clinical utility alone.

	regulatory status.

Reproducibility and accountability

Reproducibility requires more than publication of final performance values. A reproducible AI-enabled drug discovery workflow should preserve the data version, preprocessing operations, molecular representations, feature-generation procedures, model configuration, software dependencies, randomization controls, evaluation splits, statistical analyses, and reporting decisions that produced a result. Life-science machine-learning standards therefore emphasize access to sufficiently detailed data, code, computational environments, and methodological records to permit independent inspection and materially consistent rerunning of an analysis [16]. Empirical assessments of health-related machine-learning research nevertheless indicate that incomplete reporting, inaccessible code, unavailable data, and underspecified computational procedures continue to limit independent reproduction [17].

Version control and workflow traceability are necessary because even small changes in preprocessing, data inclusion, model parameters, dependencies, or evaluation design can materially alter results. Each reported output should be linked to immutable identifiers for the relevant dataset, code commit, model version, environment, and analysis run. Leakage prevention should also be treated as part of reproducibility because a workflow may be technically rerunnable while reproducing an invalid estimate produced by information crossing between development and evaluation stages [18]. Molecular model comparisons further depend on the selected representation, prediction task, data split, optimization strategy, and metric, making those choices essential components of the reproducibility record [19]. Shared benchmark resources can improve comparability by standardizing datasets, splits, and evaluation procedures, but they do not eliminate the need to disclose benchmark limitations or justify whether a task reflects the intended discovery problem [20].

Accountability concerns who is responsible for the design, validation, release, interpretation, monitoring, and use of an AI system. Responsibility should not be assigned to the model itself or obscured by references to automated decision-making. Biomedical AI governance must address opacity, bias, privacy, evidentiary limitations, and the allocation of responsibility among developers, data stewards, domain experts, institutional leaders, and decision owners [21]. Algorithmic outputs can redistribute epistemic authority by influencing which hypotheses, compounds, or experiments receive attention, while also complicating moral responsibility when users defer to a system whose operation or limitations they do not fully understand [22]. Human oversight must therefore include authority to question, reject, override, or escalate an output rather than functioning as nominal approval after a computational recommendation has already shaped the decision.

A combined reproducibility and accountability framework should require an auditable record of model runs, data access, validation decisions, deviations from protocol, expert overrides, identified risks, and corrective actions. Code and data should be made available where appropriate and lawful; when unrestricted release is not possible, controlled access, executable environments, synthetic examples, or structured inspection procedures may support verification without overriding confidentiality, privacy, security, or intellectual-property constraints. Bias assessment, fairness assessment, representativeness review, and risk analysis should be assigned to named roles and linked to predefined escalation routes. Reproducibility can strengthen evidence credibility, while accountability can clarify who must act when evidence is incomplete or a failure occurs, but neither function alone demonstrates experimental validity, clinical usefulness, regulatory acceptance, or safe implementation.

Regulatory alignment

Regulatory alignment should be understood as a structured process for organizing evidence, responsibilities, controls, and lifecycle records in a manner that can support context-specific review. It is not equivalent to regulatory approval, legal compliance, certification, or authority acceptance. This distinction is particularly important for models that may change after initial development because a one-time assessment of a locked system does not fully address continuously learning or periodically updated algorithms. Adaptive behavior creates governance questions concerning which changes are permitted, how they are documented, when revalidation is required, and who has authority to approve a revised version [23].

Experience with authorized medical AI systems illustrates that authorization status does not guarantee uniform transparency regarding development data, validation methods, subgroup performance, algorithmic changes, or post-market evidence. Comparative analyses have identified substantial variation in the publicly available information associated with AI- and machine-learning-based medical devices across regulatory settings [24]. Public cataloguing of authorized systems has likewise shown that documentation concerning datasets, algorithms, evaluation, and system characteristics may be

incomplete or inconsistently accessible [25]. Although medical-device regulation is not directly interchangeable with drug-discovery governance, these observations support a broader principle: regulatory alignment requires traceable evidence organization and should not be inferred from the presence of an AI method, a favorable benchmark, or a prior authorization in another context.

Within pharmaceutical research, alignment must account for the diversity of AI applications across target discovery, molecular design, formulation, drug delivery, translational modeling, and personalized therapeutic development. These applications differ in their data sources, evidentiary endpoints, experimental dependencies, intended users, and potential consequences, so they should not be governed through a single undifferentiated validation pathway [26]. A regulatory-science perspective on AI and machine learning in drug and biological-product development emphasizes the relevance of data quality, model credibility, documentation, validation, risk management, lifecycle evidence, and communication with appropriate oversight functions [27]. The proposed architecture therefore organizes evidence according to intended use and risk without prescribing a specific authorization route or claiming that the resulting documentation would satisfy any authority.

Regulatory-alignment records should connect the intended use, data provenance, model version, validation plan, uncertainty characterization, identified risks, monitoring indicators, update rules, responsible roles, and decision boundaries. Lifecycle monitoring should define how distributional change, calibration deterioration, unexpected errors, user overrides, and emerging evidence will be detected and assessed. Update governance should distinguish minor technical maintenance from changes that could affect model behavior, intended use, or evidentiary claims, with change-specific impact assessment and revalidation where appropriate. Post-deployment monitoring may be relevant when a system enters operational use, but it cannot substitute for adequate pre-use evaluation. A formal no-approval-claim boundary should require explicit separation between regulatory alignment, governance readiness, implementation readiness, clinical utility, legal sufficiency, and regulatory authorization [27]. **Table 2** maps reproducibility, accountability, regulatory-alignment, and implementation safeguards for AI-enabled drug discovery.

Table 2. Reproducibility, Accountability, and Regulatory-Alignment Safeguards for AI-Enabled Drug Discovery: Reproducible Workflows, Version Control, Audit Trails, Human Oversight, Bias Assessment, Lifecycle Monitoring, Model Update Governance, Evidence Documentation, and Implementation Boundaries

Governance domain	Core governance question	Required process	Responsible actor or role	Regulatory-relevance logic	Implementation barrier	Safeguard	Architecture output
Reproducible workflow	Can an independent qualified team reconstruct and rerun the analysis?	Preserve executable pipelines, preprocessing logic, model settings, seeds, environments, test cases, and access instructions.	Computational lead; reproducibility reviewer.	Supports inspection of how evidence was produced.	Hidden steps, unavailable dependencies, restricted data, or unstable environments.	Containerized environments, controlled-access resources, and independent rerun exercises.	Reproducibility package.
Version control	Which data, code, model, dependency, and document versions produced each output?	Use immutable releases, commit identifiers, change logs, approval records, and rollback procedures.	Model owner; configuration manager.	Preserves the relationship between evaluated and operational versions.	Silent changes and dependency drift.	Release controls and automatic propagation of version identifiers into outputs.	Versioned model record.
Model update tracking	What changed, why did it	Classify the change, perform	Change-control board;	Supports controlled lifecycle	Frequent retraining, changing data	Locked production versions,	Approved change package.

	change, and what evidence supports the revised version?	impact analysis, update documentation, repeat applicable validation, and obtain approval.	model owner; validation lead.	modification.	sources, and unclear materiality thresholds.	change categories, rollback capability, and revalidation rules.	
Audit trail	Can evidence, actions, and decisions be reconstructed after a challenge or incident?	Record model runs, data access, validation decisions, approvals, overrides, deviations, and corrective actions.	Quality function; system owner; audit reviewer.	Enables inspection and incident analysis.	Fragmented systems, mutable logs, and inconsistent retention.	Tamper-evident, access-controlled, time-stamped records.	Auditable event ledger.
Human oversight	Which outputs require expert review, challenge, rejection, or escalation?	Define reviewer competence, review points, override authority, escalation triggers, and documentation duties.	Domain expert; accountable decision owner.	Demonstrates controlled use of consequential outputs.	Automation bias, workload, and nominal rather than substantive review.	Mandatory review for high-risk, high-uncertainty, or out-of-domain outputs.	Human-oversight plan.
Decision accountability	Who owns each model-related decision and its consequences?	Assign responsibility for development, validation, release, use, monitoring, and corrective action.	Governance board; scientific or business decision owner.	Clarifies who is answerable across the evidence chain.	Diffused responsibility across teams, institutions, or vendors.	Responsibility matrix and signed decision records.	Accountability map.
Bias assessment	Which systematic errors may arise from sampling, labels, assays, modeling, or use?	Review data generation, chemical-series composition, subgroup behavior, assay bias, missingness, and downstream effects.	Data steward; validation lead; domain specialist.	Bias findings inform risk, validity, and use restrictions.	Sparse coverage, unavailable protected attributes, and hidden historical selection effects.	Bias register, sensitivity analysis, and escalation thresholds.	Bias assessment report.
Fairness assessment	Are errors, benefits, or burdens distributed appropriatel	Define the fairness question, evaluate suitable	Fairness reviewer; affected-domain representativ	Shows that equity-related risks were considered	Conflicting fairness definitions and limited	Transparent normative rationale and documente	Fairness assessment statement.

	y across relevant groups or contexts?	measures, document trade-offs, and assess consequences.	es; governance board.	without asserting universal fairness.	contextual data.	d unresolved trade-offs.	
Representativeness review	Does the evidence cover the intended population, chemical space, assay domain, and operational context?	Compare development evidence with intended-use distributions and document coverage gaps.	Data steward; domain expert; validation lead.	Supports defensible intended-use boundaries.	Convenience sampling and narrow historical datasets.	Coverage analysis, applicability-domain definition, and explicit exclusions.	Representativeness statement.
Risk management	What harms or scientific failures could arise, and how will they be controlled?	Identify hazards, estimate likelihood and impact, assign controls, evaluate residual risk, and maintain a risk register.	Risk owner; cross-functional governance board.	Organizes oversight according to consequence and uncertainty.	Poorly defined intended use and unknown failure modes.	Periodic risk review linked to model, data, and context changes.	AI risk-management file.
Lifecycle monitoring	What evidence indicates that the system remains suitable over time?	Monitor input distributions, missingness, calibration, errors, overrides, incidents, and operational context.	Monitoring team; model owner; quality function.	Provides continuing evidence after implementation where relevant.	Delayed outcomes, limited ground truth, and weak monitoring infrastructure.	Predefined indicators, thresholds, review intervals, and escalation routes.	Lifecycle-monitoring plan.
Model drift detection	How will distributional, calibration, or performance change be detected?	Establish baselines and test covariate, concept, calibration, and performance drift.	Monitoring team; validation lead.	Supports timely reassessment of continued suitability.	Low event frequency and uncertain alert significance.	Multi-signal monitoring followed by expert investigation.	Drift assessment record.
Evidence documentation	Is each claim linked to reviewable evidence and limitations?	Connect claims to datasets, analyses, validation results, uncertainties, approvals,	Evidence curator; validation lead; decision owner.	Creates a structured basis for context-specific review.	Documentation burden and inconsistent terminology.	Controlled templates, evidence indexes, and completeness checks.	Evidence dossier.

		and unresolved issues.					
Regulatory alignment	Which requirements, standards, or review questions may be relevant to the intended use?	Maintain a context-specific mapping between architecture artifacts and applicable expectations.	Regulatory-science lead; qualified legal or compliance advisers where appropriate.	Supports governance readiness without implying conformity or approval.	Changing requirements and jurisdictional variation.	Living alignment matrix and periodic specialist review.	Regulatory-alignment map.
Implementation safeguards	What prevents unsupported, unsafe, or out-of-scope use?	Define access controls, user competence, use restrictions, abstention behavior, manual checks, and incident procedures.	Implementation lead; system owner; domain expert.	Demonstrates control of operational context.	Workarounds, misuse, and workflow mismatch.	Role-based access, training, alerts, and compliance review.	Safeguard specification.
Post-deployment monitoring where relevant	What evidence should be collected after operational introduction?	Track incidents, near misses, performance, use patterns, overrides, and unintended effects.	Operational owner; safety or quality function.	Supports continuing oversight without replacing pre-use evidence.	Under-reporting and ambiguous event definitions.	Standard reporting routes and periodic safety review.	Post-use evidence report.
No approval-claim boundary	How will alignment be distinguished from authorization, certification, legal sufficiency, and clinical utility?	Review terminology in documentation, publications, dashboards, and decision records.	Senior accountable owner; regulatory-science reviewer; editorial lead.	Prevents governance artifacts from being misrepresented as authority endorsement.	Promotional pressure and ambiguous language.	Prohibited-claim checklist and pre-release language review.	Formal boundary statement.

Proposed trust architecture

The proposed trust architecture is a conceptual, layered governance structure connecting AI-enabled drug discovery activities with the evidence and oversight required to support trustworthy use. Its first layer defines the scientific workflow, including data collection, molecular representation, model development, candidate prediction, experimental-validation planning, and bounded decision support. These activities are surrounded by cross-cutting controls rather than treated as a self-governing technical pipeline. The accountability layer requires fairness questions to be defined in relation to the relevant populations, scientific outcomes, error distributions, and

consequences, recognizing that no single fairness measure is appropriate for every context [28].

The transparency layer links every model output to documented data provenance, dataset characteristics, model assumptions, software configurations, explanation methods, applicability conditions, and uncertainty. Representativeness is assessed by comparing the chemical, biological, experimental, and population coverage of the development evidence with the intended context. Inadequate or historically skewed data can reproduce existing disparities or create systematic blind spots, making coverage analysis and explicit exclusion boundaries essential components of the architecture [29].

The validation layer then evaluates only the claims defined in the validation plan through internal testing, independent external evidence, prospective assessment where relevant, and experimental confirmation where relevant.

The reproducibility layer preserves the computational and methodological conditions necessary to reconstruct the evidence, while the lifecycle layer addresses changes that arise when data sources, assay practices, compound distributions, disease contexts, users, or operational environments evolve. Dataset shift should be anticipated before implementation rather than treated solely as an unexpected post-use event [30]. The architecture therefore requires baseline characterization, shift indicators, applicability-domain checks, versioned monitoring rules, and predefined responses such as investigation, restricted use, recalibration, revalidation, rollback, or temporary suspension of the affected function.

The accountability and regulatory-alignment layers assign named roles for data stewardship, model development, independent validation, scientific review, risk ownership, release approval, monitoring, and corrective action. Auditability connects these roles to time-stamped evidence and decisions. Lifecycle monitoring should include calibration because model predictions may become progressively misaligned with observed outcomes even when the computational code remains unchanged [31]. Human oversight operates across all layers through substantive review rights, authority to reject or override outputs, and escalation requirements for high uncertainty, unsupported domains, validation failures, unexplained drift, or unresolved bias.

The final layer is a controlled feedback and governance loop. New experimental evidence, failed predictions, audit findings, user feedback, monitoring signals, risk reviews, and regulatory-science developments may trigger revision of the dataset, model, validation plan, safeguards, intended-use statement, or architecture documentation. Proposed changes must return through applicable transparency, validation, reproducibility, accountability, and approval controls rather than being introduced as undocumented adaptation. Decision boundaries explicitly separate prediction from evidence, explanation from validity, reproducibility from utility, governance readiness from implementation readiness, and regulatory alignment from regulatory approval. **Table 3** presents the proposed trust architecture for AI-enabled drug discovery. **Figure 1** illustrates the proposed trust architecture for AI-enabled drug discovery, connecting transparency, validation, reproducibility, accountability, regulatory alignment, lifecycle monitoring, human oversight, and feedback governance.

Table 3. Proposed Trust Architecture for AI-Enabled Drug Discovery: Transparency, Validation, Reproducibility, Accountability, Bias Control, Auditability, Lifecycle Monitoring, Regulatory Alignment, Human Oversight, Feedback, and Decision Boundaries

Architecture layer	Core purpose	Required input	Governance or technical function	Potential output	Failure risk	Oversight requirement	Feedback mechanism	Decision boundary
AI-enabled drug discovery workflow	Define the scientific activities and decisions governed by the architecture.	Molecular, biological, experimental, clinical, and contextual data; intended-use statement.	Organizes data collection, representation, model development, candidate prediction, experimental planning, and bounded decision support.	Prioritized hypotheses, compounds, targets, experiments, or evidence summaries.	Technical outputs used beyond their supported scientific purpose.	Domain-expert review and named decision ownership.	Experimental findings and downstream decision outcomes return to the evidence record.	Computational output is advisory evidence, not proof of biological validity, safety, efficacy, or approval.
Transparency	Make data, models, assumptions,	Provenance records, dataset documentation	Provides traceability, interpretability, explanation, and	Dataset record, model record, traceable output,	Hidden assumptions, misleading explanations,	Data steward, model owner, and	Documentation is updated when data, models,	Transparency may support trustworthiness but cannot

	limitations, and uncertainty inspectable.	, model documentation , explanation methods, version identifiers.	uncertainty disclosure.	explanation report, uncertainty statement.	or irreconstructable evidence.	independent reviewer.	assumptions, or intended use change.	establish validity alone.
Validation	Test clearly defined claims in relevant contexts.	Validation plan, independent datasets, benchmarks, comparators, endpoints, acceptance criteria.	Conducts internal, external, prospective, and experimental validation where relevant.	Validation report, calibration analysis, failure analysis, applicability-domain statement.	Overfitting, leakage, invalid benchmarks, or unsupported generalization.	Independent validation lead and domain specialist.	Validation failures trigger model revision, claim restriction, or additional evidence generation.	Validation supports only the tested claim, population, domain, and context.
Reproducibility	Preserve the conditions needed to reconstruct and independently inspect results.	Versioned data, code, environments, preprocessing logic, model configuration, run records.	Enables rerunning, replication attempts, and verification of the computational evidence chain.	Reproducibility package and reconstruction report.	Hidden steps, unavailable resources, unstable environments, or irreproducible outputs.	Reproducibility reviewer and configuration manager.	Failed reconstruction triggers documentation correction or workflow remediation.	Reproducibility does not establish biological correctness, utility, or approval.
Accountability and bias control	Assign responsibility and examine unequal or systematic error.	Responsibility map, risk register, bias evidence, fairness rationale, representativeness analysis.	Allocates ownership, assesses bias and fairness, and links risks to controls.	Accountability map, bias report, fairness statement, representativeness statement.	Diffused responsibility, unrecognized bias, inequitable effects, or unresolved risk.	Governance board, risk owner, fairness reviewer, and affected-domain experts.	Incidents and bias findings trigger escalation, corrective action, and reassessment.	No output proceeds when responsibility is unassigned or material bias remains unmanaged.
Auditability and traceability	Preserve a reconstructable record of evidence, actions, and decisions.	Run identifiers, access logs, validation records, approvals, overrides, deviations, and change history.	Creates a time-stamped linkage between evidence and responsible action.	Auditable event ledger and decision record.	Inability to investigate errors, disputes, or unauthorized changes.	Quality or audit function with appropriate independence .	Audit findings enter the corrective-action and governance-update process.	Untraceable or materially incomplete decisions cannot support consequential progression.
Regulatory alignment	Organize evidence and controls for context-specific governance review.	Intended use, evidence dossier, risk file, monitoring plan, update rules, safeguard specification.	Maps architecture artifacts to relevant review questions without asserting conformity.	Regulatory-alignment map and governance-readiness record.	Alignment language misrepresented as approval or legal sufficiency.	Regulatory-science lead and qualified specialist review.	Changes in requirements or intended use trigger remapping and evidence review.	Regulatory alignment is not regulatory approval, certification, or authority acceptance.
Lifecycle monitoring and update governance	Detect change and control modifications after initial validation.	Baselines, monitoring indicators, drift thresholds, incident reports, update proposals.	Detects distributional, calibration, performance, and contextual drift; controls updates.	Drift assessment, monitoring report, approved change package,	Silent degradation, uncontrolled retraining, or continued use after evidence becomes outdated.	Monitoring team, change-control board, model owner, and validation lead.	Monitoring signals trigger investigation, recalibration, revalidation, restriction,	Material changes require impact assessment and applicable revalidation

				rollback decision.			rollback, or suspension.	before release.
Human oversight	Ensure consequential decisions remain reviewable and contestable.	Output, uncertainty, explanation, validation status, risk level, and intended-use boundary.	Enables expert review, challenge, override, abstention, and escalation.	Documented expert decision, override, rejection, or request for further evidence.	Automation bias, nominal review, or inappropriate deference to the model.	Qualified domain expert with explicit authority and sufficient information.	Reviewer decisions and concerns enter monitoring, audit, and feedback records.	High-risk, uncertain, out-of-domain, or failed-validation outputs require escalation or abstention.
Feedback governance and decision boundaries	Convert new evidence and failures into controlled improvement while limiting unsupported claims.	Experimental results, audit findings, incidents, user feedback, monitoring signals, regulatory-science developments.	Triages feedback, prioritizes changes, updates documentation, and routes revisions through governance gates.	Corrective-action log, revised risk controls, updated validation plan, restricted intended use, or revised model.	Uncontrolled adaptation, confirmation bias, unresolved feedback, or claim escalation.	Governance board and accountable architecture owner.	Closed-loop review links feedback to documented action and follow-up verification.	Feedback does not authorize uncontrolled model change, and governance readiness does not equal implementation or approval readiness.

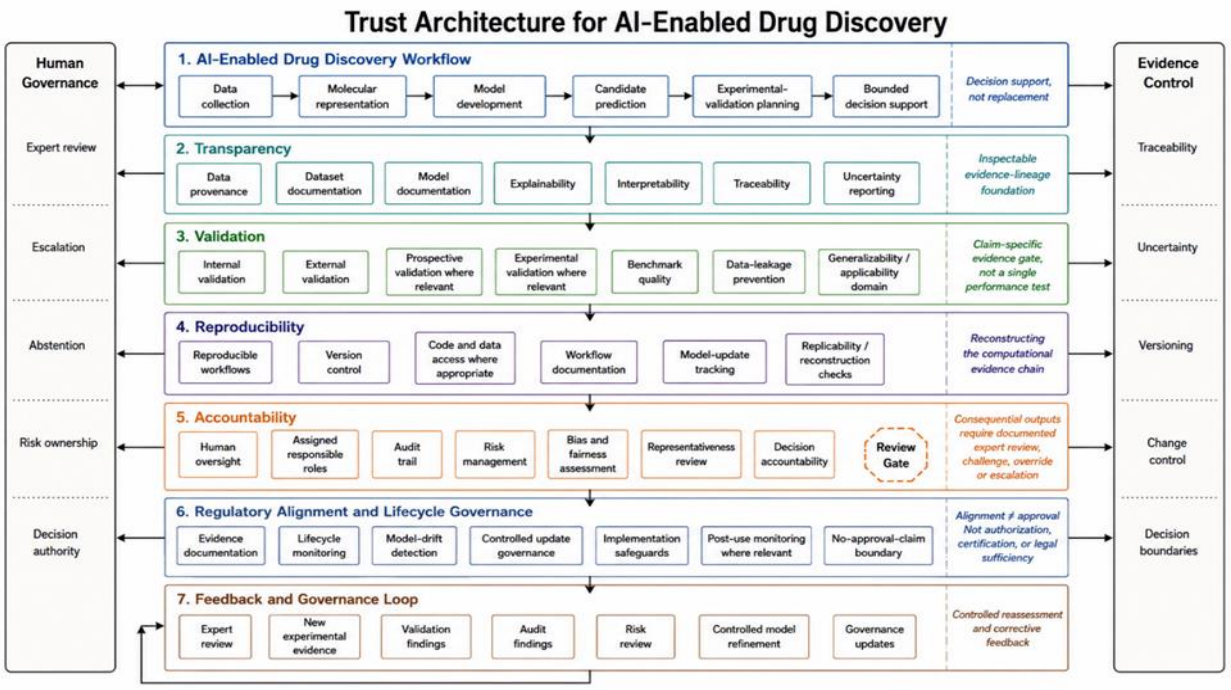

Figure 1. Trust Architecture for AI-Enabled Drug Discovery

Alt-text description

A conceptual governance architecture showing AI-enabled drug discovery workflows supported by transparency, validation, reproducibility, accountability, bias control, audit trails, lifecycle monitoring, regulatory alignment, human oversight, and feedback loops.

Implementation implications

Implementation requires translation of broad trustworthy-AI principles into specific institutional roles, evidence artifacts, review procedures, and maintenance responsibilities. Ethical and governance scholarship in health AI has repeatedly identified a gap between high-level principles and operational practice [32]. Drug discovery organizations should therefore assign ownership for data stewardship, model documentation, validation planning, reproducibility review, experimental coordination, bias assessment, release decisions, monitoring, and corrective action. Governance boards should review the intended use, evidence boundaries, unresolved uncertainties, and escalation rules rather than focusing only on aggregate performance.

Computational teams should prepare reproducibility packages and leakage analyses before consequential model use, while experimental collaborators should participate in defining biologically meaningful endpoints and the conditions under which computational predictions require laboratory confirmation. Quality and governance functions should maintain responsibility maps, risk registers, audit trails, and model-change records. Bias governance should be treated as an ongoing safety and quality-improvement responsibility rather than a one-time statistical exercise, with new evidence and identified disparities feeding into reassessment and corrective action [33]. Journal reviewers and funders can support these practices by requesting sufficient documentation to evaluate provenance, validation design, benchmark credibility, reproducibility, limitations, and claim boundaries.

The architecture also requires maintenance. Monitoring plans, drift thresholds, responsible roles, access arrangements, and validation records may become outdated as data, software, assays, models, users, and scientific questions change. Periodic architecture review should therefore determine whether the intended-use statement remains appropriate, whether documentation is current, whether monitoring remains feasible, and whether new evidence changes the risk profile. These implementation implications are research-governance recommendations rather than legal, regulatory, medical, or clinical advice. Applying them may strengthen auditability and governance readiness, but it does not establish compliance, authorization, clinical utility, or deployment readiness.

CONCLUSION

Trustworthy AI-enabled drug discovery requires an integrated evidence and governance structure rather than reliance on model performance alone. The proposed trust architecture connects data provenance, dataset and model documentation, explainability, interpretability, traceability, validation, reproducibility, accountability, bias control, auditability, regulatory alignment, lifecycle monitoring, human oversight, controlled updating, and feedback governance. It also establishes decision boundaries that prevent transparency from being treated as proof of validity, reproducibility as proof of usefulness, or regulatory alignment as proof of approval. By assigning responsibilities and linking scientific claims to documented evidence, monitoring, and escalation processes, the architecture may strengthen the credibility and reviewability of AI-supported discovery decisions. It remains a conceptual governance contribution that requires context-specific adaptation, evaluation, and expert oversight before operational use.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[2] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009

[3] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037

[4] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010

[5] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-15. doi:10.1038/s42256-019-0048-x

[6] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2

[7] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

[8] Sieg J, Flachsenberg F, Rarey M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712

[9] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723

[10] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y

[11] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310. doi:10.1186/s12911-020-01332-6

[12] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50. doi:10.1016/S2589-7500(21)00208-9

[13] Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Harrow J, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4

[14] Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018;286(3):800-9. doi:10.1148/radiol.2017171920

[15] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

[16] Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7

[17] McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med. 2021;13(586):eabb1655. doi:10.1126/scitranslmed.abb1655

[18] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804

[19] Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237

[20] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A

[21] Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med. 2018;15(11):e1002689. doi:10.1371/journal.pmed.1002689

[22] Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586

[23] Babic B, Gerke S, Evgeniou T, Cohen IG. Algorithms on regulatory lockdown in medicine. Science. 2019;366(6470):1202-4. doi:10.1126/science.aay9547

[24] Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. Lancet Digit Health. 2021;3(3):e195-203. doi:10.1016/S2589-7500(20)30292-2

[25] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. NPJ Digit Med. 2020;3:118. doi:10.1038/s41746-020-00324-0

[26] Serrano DR, Luciano FC, Anaya BJ, Ongoren B, Kara A, Molina G, et al. Artificial intelligence applications in drug discovery and drug delivery: revolutionizing personalized medicine. Pharmaceutics. 2024;16(10):1328. doi:10.3390/pharmaceutics16101328

[27] Mirakhori F, Niazi SK. Harnessing the AI/ML in drug and biological products discovery and development: the regulatory perspective. Pharmaceuticals (Basel). 2025;18(1):47. doi:10.3390/ph18010047

[28] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-72. doi:10.7326/M18-1990

[29] Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med. 2020;26(1):16-7. doi:10.1038/s41591-019-0649-2

[30] Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable

models in health AI. Biostatistics. 2020;21(2):345-52. doi:10.1093/biostatistics/kxz041

[31] Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61. doi:10.1093/jamia/ocx030

[32] Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med. 2020;260:113172. doi:10.1016/j.socscimed.2020.113172

[33] McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Zlotnik Shaul R. Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. J Am Med Inform Assoc. 2020;27(12):2024-7. doi:10.1093/jamia/ocaa085