



Paper Type: Original Article

Managing Reproducibility Risk in Multi-Omics Analytics

Niloofer Jamareh^{1,*}, Neda Jamareh²

¹ Department of Biosciences, Università degli Studi di Milano, Via Celoria 26, 20133 Milan, Italy; niloofer.jamareh@studenti.unimi.it.

² Department of Mathematics and Physics, University of Campania “Luigi Vanvitelli”, Viale Lincoln 5, 81100 Caserta, Italy; neda.jamareh@studenti.unicampania.it.

Citation:

Received: 13 March 2025

Revised: 16 November 2026

Accepted: 22 October 2026

Jamareh, N., & Jamareh, N. (2026). Managing reproducibility risk in multi-omics analytics. *Management Analytics and Social Insights*, 3(4), 302–316.

Abstract

Multi-omics analytics can integrate genomic, transcriptomic, epigenomic, proteomic, metabolomic, and related data into richer representations of biological systems, but analytical flexibility can also make strong results sensitive to defensible changes in preprocessing, integration, validation, or context. This paper examines that problem as reproducibility risk at the decision level. We combine a structured narrative synthesis of multi-omics methodology, computational reproducibility, biomedical machine learning governance, and organizational decision research with an illustrative Monte Carlo stress test. We define the Analytical Fragility Index (AFI) as the proportion of decision outputs that change across pre-specified plausible perturbations. In 500 simulated experiments with three omics signal layers, increasing technical confounding made an internally selected pipeline look progressively stronger while becoming less transportable. At the highest confounding level, mean development Area Under The Receiver Operating Characteristic Curve (AUROC) reached 0.978, replication AUROC fell to 0.753, the optimism gap reached 0.225, and 29.8% of binary decisions changed after technical context shift. These are illustrative rather than empirical estimates. We therefore propose six risk-based evidence gates linking provenance, design alignment, validation, stress testing, independent challenge, and post-decision learning. The implication is that multi-omics programs should manage analytical quality by the stability and consequence of decisions, not by model accuracy alone.

Keywords: Multi-omics analytics, Reproducibility risk, Data governance, Monte Carlo simulation, Analytical fragility, Decision management.

1 | Introduction

1.1 | Motivation and Research Problem

High-throughput bioscience increasingly studies biological systems through several molecular layers rather than one assay at a time. Genomics, transcriptomics, epigenomics, proteomics, metabolomics, and phenotypic information can be integrated to identify mechanisms, molecular subtypes, biomarkers, and treatment-response patterns that are difficult to recover from a single data source [1], [2]. The frontier is moving further

Corresponding Author: niloofer.jamareh@studenti.unimi.it



<https://doi.org/10.22105/masi.v3i4.115>



Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

toward Artificial Intelligence (AI)-assisted multimodal integration and foundation-model approaches that combine molecular and clinical information at scale [3], [4].

This expansion creates a methodological opportunity and a management problem. Multi-omics analysis does not offer one universally dominant pipeline: Integration can occur early, at latent-representation stages, or after modality-specific modeling, and each strategy makes different assumptions about scale, missingness, dependence, and biological alignment [5-7]. Brooks et al. [8] show why omics benchmarking itself requires careful design because conclusions can depend on datasets, metrics, ground truth, and implementation choices. Recent benchmarks illustrate the size of the choice set: Hu et al. [9] compared 14 cross-modality prediction algorithms and 18 integration algorithms using 47 single-cell multi-omics datasets, while Fu et al. [10] evaluated 40 single-cell multimodal integration algorithms across paired, unpaired, and mosaic settings.

For a bioscience laboratory, research center, biotechnology program, or translational team, the decision is not merely which model produces the highest score. Analytical outputs determine which biomarker is advanced, which pathway receives experimental follow-up, which assay is repeated, which candidate enters validation, and which project consumes scarce samples, instrument time, personnel, and budget. If a conclusion changes under a reasonable alternative normalization, integration method, train-test partition, batch treatment, or replication context, the organization faces a reproducibility risk even when the original model looks statistically impressive.

We define reproducibility risk as the probability and consequence of a decision-relevant conclusion failing to remain stable when the same scientific question is evaluated under analytically plausible or operationally relevant changes. This definition deliberately goes beyond exact computational rerunability. A workflow can be perfectly rerunnable yet remain fragile to an alternative but defensible pipeline. Conversely, some changes in results are biologically informative rather than problematic. The practical challenge is to distinguish useful biological heterogeneity from avoidable analytical dependence and to allocate review effort according to the consequence of getting the decision wrong.

1.2 | Research Gap, Innovation, and Contributions

Three gaps motivate the paper. First, multi-omics guidance is rich on algorithm selection and benchmark performance but less explicit about how analytical uncertainty should alter decision rights. Second, reproducibility initiatives emphasize provenance, code, reporting, and leakage control, yet reproducibility is often treated as a property of the workflow rather than the stability of the downstream scientific decision. Third, biomedical AI governance has advanced rapidly, but much of it targets clinical prediction and deployment rather than research-stage portfolio choices such as assay prioritization, biomarker advancement, and experimental commitment.

The paper makes four contributions. First, it formalizes decision-level analytical fragility and proposes an Analytical Fragility Index (AFI) that can be calculated from a pre-specified set of plausible pipeline perturbations. Second, it provides an original Monte Carlo stress test showing how internal predictive performance can rise while replication performance and decision stability deteriorate when technical structure is spuriously informative. Third, it develops a six-gate, risk-based governance architecture that links evidence quality to escalation, replication, deferral, and stop decisions. Fourth, it translates the framework into testable propositions and measurable operational outcomes, including decision flip rates, optimism gaps, rework, time-to-reproduce, external validation loss, and recurrence of previously diagnosed failure modes.

A counterargument is that adding governance around exploratory bioscience can slow discovery and discourage productive methodological experimentation. That risk is real. The framework therefore does not recommend universal review. Its central design principle is proportionality: Low-consequence, reversible exploration should remain fast, whereas high-consequence decisions with high analytical fragility should trigger stronger challenge, replication, or deferral.

1.3 | Main Work and Paper Organization

Section 2 reviews the evidence on multi-omics integration, benchmarking, reproducibility, leakage, documentation, and translational governance. Section 3 introduces quantitative measures of analytical fragility and reports the illustrative Monte Carlo stress test. Section 4 converts those measures into a risk-based governance architecture. Section 5 develops managerial implications, propositions, maturity levels, and empirical research designs. Section 6 concludes with limitations and future work. The numerical experiment uses only synthetic data and is intended as a mechanism demonstration rather than an estimate of the prevalence or magnitude of reproducibility failures in real laboratories.

2 | Literature Review

2.1 | Multi-Omics Integration and Method-Dependent Evidence

The central promise of multi-omics is to integrate complementary biological information while respecting the different statistical structures of each modality. Ritchie et al. [5] distinguish several ways to combine data for genotype-phenotype analysis and emphasize that integration choices must follow the biological question. Picard et al. [6] similarly organize machine-learning strategies by integration stage, while Acharya and Mukhopadhyay review how machine learning is used to fuse genomic, transcriptomic, proteomic, metabolomic, and clinical information in precision oncology [7].

The software ecosystem makes sophisticated integration broadly accessible. Mixomics implements multivariate feature selection and supervised or unsupervised integration [11]. Multi-Omics Factor Analysis (MOFA) models shared and view-specific latent factors across omics layers [12], and MOFA+ extends this framework to larger and more complex multimodal designs [13]. These tools are valuable, but accessibility increases the number of analytically reasonable choices that can be made after seeing the data. A result can depend on feature filtering, normalization, latent dimensionality, modality weighting, missing-data strategy, hyperparameter selection, or the benchmark metric chosen to define success.

Benchmarking studies show why this matters. No single integration method dominates every combination of modality, task, dataset size, and quality [8-10]. The implication is not that benchmarking is unreliable, but that the evidence supporting a scientific decision is conditional on the benchmark design. For management, this means method selection should be treated as an assumption-bearing decision rather than an invisible technical step.

2.2 | Batch Effects, Preprocessing, and Data Leakage

Analytical fragility can enter before modeling. RNA sequencing workflows depend on experimental design, quality control, alignment, quantification, normalization, and downstream testing choices [14]. Batch effects can create systematic differences that mimic biological signal or hide real effects [15]. In single-cell studies, confounding between technical batches and biological groups can make technical and biological variation difficult or impossible to separate after data collection [16]. Cross-modality best-practice guidance therefore emphasizes experimental design, quality control, batch-aware analysis, and validation across modalities [17].

Machine learning introduces an additional threat: information can leak from the evaluation set into model development through feature selection, preprocessing, repeated tuning, related samples, or dependency structures that violate the intended prediction scenario. Barnett et al. [18] provide seven diagnostic questions specifically for biological machine-learning applications and emphasize that leakage can be subtle because biological observations are often dependent. Kapoor and Narayanan [19] show more broadly how leakage can generate overoptimistic scientific claims and contribute to a reproducibility crisis in machine-learning-based research. These problems are especially relevant to multi-omics because the feature space is large, samples are often limited, and the same subjects, batches, sites, or biological hierarchies can appear across several data views.

2.3 | Reproducibility, Provenance, and Documentation

The Findable, Accessible, Interoperable, and Reusable (FAIR) principles frame data stewardship around discoverability, access, interoperability, reuse, and provenance [20]. Computational reproducibility guidance adds operational practices: Preserve code, avoid undocumented manual steps, record software environments, link raw and processed data, and keep intermediate outputs when they are needed to reconstruct results [21]. Reproducibility is therefore not a single technical feature; it is a system of data, code, workflow, metadata, and organizational practice. Munafò et al. [22] argue accordingly that improving reproducible science requires changes in methods, reporting, incentives, and institutional processes.

The biomedical consequences of weak reproducibility are not merely academic. Begley and Ellis [23] highlighted failures to reproduce influential preclinical cancer findings, while Freedman et al. [24] estimated large economic costs from irreproducible preclinical research. The national institutes of health subsequently emphasized rigor and transparency as policy priorities [25]. In health machine learning, McDermott et al. [26] documented continuing gaps in reproducibility, and Roberts et al. [27] showed how methodological weaknesses in COVID-19 imaging models could prevent apparently strong models from supporting clinical use.

Documentation tools make analytical assets easier to govern. Datasheets for datasets formalize data purpose, collection, composition, and limitations [28], model cards do the same for model purpose, evaluation, and constraints [29]. Sculley et al. [30] show that machine-learning systems also accumulate hidden technical debt through data dependencies, entanglement, feedback, and configuration complexity. In a multi-omics program, equivalent documentation can make provenance, exclusions, assay versions, integration decisions, and model changes visible to reviewers who were not present during the original analysis.

2.4 | From Reproducibility to Decision Governance

Reproducibility alone does not guarantee transportability. Clinical and biomedical machine learning systems can fail when patient populations, measurement processes, laboratories, or workflows differ from development conditions [31], [32]. Responsible machine-learning roadmaps therefore emphasize clinically meaningful validation, human factors, monitoring, and the consequences of error [33], [34]. Reporting and risk-of-bias standards have also become more specific: The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence (TRIPOD+AI) guideline updates reporting expectations for prediction models [35], and the Prediction model Risk of Bias ASsessment Tool plus Artificial Intelligence (PROBAST+AI) addresses quality, bias, and applicability [36]. The Future-AI consensus extends trustworthy AI across the lifecycle through fairness, universality, traceability, usability, robustness, and explainability [37].

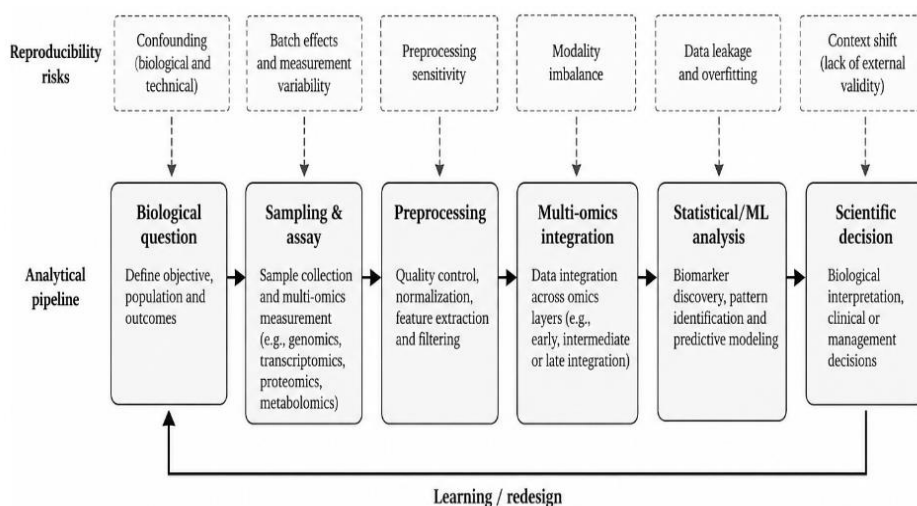
Broader governance frameworks reinforce the same lifecycle logic. The world health organization emphasizes accountability, transparency, inclusiveness, safety, and sustainability in AI for health [38]. The National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (AI RMF) organizes risk management around governance, mapping, measurement, and management across the AI lifecycle [39]. These frameworks are valuable, but they do not directly answer a recurring research-management question: When should a bioscience team be allowed to proceed with a fragile but exploratory result, and when should the same analytical fragility block or defer a costly downstream commitment?

Management theory helps articulate that missing layer. Exploration-exploitation theory explains why organizations must balance experimentation with disciplined commitment [40]. Organizational learning research shows that experience becomes useful only when it is encoded, retained, and reused [41]. Work on managing AI and algorithmic decision structures further shows that the allocation of authority, accountability, and human challenge can matter as much as technical capability [42-44]. The unresolved gap is therefore not another universal checklist. It is a rule that connects analytical fragility and decision consequence to evidence requirements, escalation, and learning.

Table 1. Literature streams and the unresolved decision-governance gap.

Literature Stream	What it Contributes	What Remains Unresolved
Multi-omics integration and benchmarking	Integrates molecular layers; benchmarks task- and dataset-specific performance.	Strong benchmark performance does not establish downstream decision stability.
Reproducibility and provenance	Provides data/code lineage, rerunnable workflows, and transparent documentation.	Exact rerunning can faithfully reproduce a fragile result.
Biomedical ML validation and governance	Adds leakage control, reporting, bias/applicability review, and monitoring.	Guidance is mainly clinical or generic and weakly linked to research-stage decision rights.
Management and organizational learning	Clarifies decision rights, proportional oversight, and feedback learning.	Lacks bioscience-specific fragility signals and analytical evidence gates.

Fig. 1 integrates these literatures at the level of the analytical pipeline. Reproducibility risks can originate at study design and measurement, not only at the final model, and downstream sophistication can make an upstream problem harder to notice. The figure also emphasizes feedback: Post-decision evidence should modify future sampling, analysis, and validation rather than remain isolated in project archives.

**Fig. 1. Multi-omics analytical pipeline and major sources of reproducibility risk.**

3 | Measuring Reproducibility Risk

3.1 | Decision-Level Analytical Fragility

Let a baseline analysis produce a decision vector $\hat{d}^{(0)}$ for n decision units, where a unit can be a sample classification, biomarker advancement decision, pathway shortlist, or any other output that drives action. Let $k = 1, \dots, K$ index pre-specified, scientifically plausible perturbations to preprocessing, normalization, missing-data handling, integration, model specification, validation, or deployment context. A simple decision-level AFI is the average proportion of decisions that change relative to the baseline:

$$AFI = (1 / nK) \sum_{i=1}^n \sum_{k=1}^K I[d_i^{(0)} \neq d_i^{(k)}]. \quad (1)$$

AFI ranges from 0 to 1. A value near 0 indicates that the action-relevant conclusion is stable across the chosen perturbation set; a high value indicates that the decision depends strongly on analytically reasonable choices. The perturbation set must be specified before inspecting the desired result, otherwise robustness analysis can itself become selective. For ranked biomarker or pathway lists, the indicator in Eq. (1) can be replaced by top-k Jaccard overlap, rank correlation, or another task-specific stability statistic. Predictive performance should be stress-tested separately. If M_{internal} is the performance estimated in the development workflow and

M_{stress} is the corresponding performance under blocked, external, or context-shifted evaluation, the optimism gap is:

$$O = M_{\text{internal}} - M_{\text{stress}}. \quad (2)$$

For area-under-the-curve metrics, a positive O indicates optimistic internal performance. AFI and O answer different questions: A model can retain similar discrimination while changing which samples cross a decision threshold, and it can retain stable classifications while its calibrated risk estimates deteriorate. Governance should therefore avoid reducing reproducibility to one metric.

Table 2. Quantitative indicators for reproducibility-risk management.

Indicator	Definition or Example	Management Interpretation
AFI	Average decision flip rate across pre-specified plausible pipeline or context perturbations.	Directly measures whether an action-relevant conclusion survives reasonable alternatives.
Optimism gap (O)	Internal performance minus blocked, external, or stress-test performance.	Detects performance that depends on development-specific structure.
Top-k stability	Jaccard overlap or rank correlation for biomarkers, pathways, or candidate lists.	Useful when the decision is a shortlist rather than a binary classification.
Modality dependence	Change in result after modality ablation, reweighting, or missing-view analysis.	Shows whether an integrated conclusion is dominated by one data source or technical layer.
Reconstruction time	Time required for an independent analyst to rebuild the decision package from recorded assets.	Operational measure of provenance completeness and handoff quality.
External validation loss	Change in discrimination, calibration, or utility in a new site, laboratory, cohort, or assay context.	Measures transportability and determines whether a result is ready for broader use.

3.2 | Illustrative Monte Carlo Stress Test

To demonstrate the mechanism without fabricating empirical bioscience results, we designed a stylized Monte Carlo experiment. The simulation does not claim to reproduce any specific omics platform. Its purpose is to isolate one common failure mode: A technical factor that becomes associated with the outcome in development data and is therefore rewarded by internal model selection, even though that association does not persist when the technical context changes. Each Monte Carlo run contains $n = 500$ binary outcomes and three omics signal layers. In each layer, class membership shifts the mean biological signal by ± 0.45 while standard-normal noise is added. The three layer-specific scores are averaged and standardized. In development data, the technical factor has a class-dependent mean of $\pm \gamma$; in the replication context, the technical factor is standard normal and independent of class. Six candidate pipelines use different weights on the technical factor, so internal model selection can favor a shortcut that disappears after technical context shift.

$$S_i(w) = Z_i + wT_i, \quad w \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}. \quad (3)$$

The technical-confounding coefficient γ takes values 0, 0.4, 0.8, and 1.2. Within each run, the pipeline weight w that maximizes development Area Under the Receiver Operating Characteristic Curve (AUROC) is selected and then applied unchanged to the replication context. Binary decisions use the zero threshold. We

repeat the procedure 500 times per confounding level and report means with empirical 2.5th and 97.5th percentile Monte Carlo intervals. Because the same biological signal is preserved while the technical association disappears, any increasing internal-replication gap is attributable to reliance on development-specific technical structure by construction.

3.3 | Simulation Results

The simulation produces a clear separation between internal performance and reproducibility as technical confounding strengthens. With $\gamma = 0$, the selected pipeline assigns essentially no weight to the technical factor (mean $w = 0.010$), mean development AUROC is 0.864, replication AUROC is 0.864 after rounding, and only 0.4% of binary decisions change. When γ increases to 0.8, internal selection shifts toward the technical factor (mean $w = 0.827$): Development AUROC rises to 0.944, but replication AUROC falls to 0.778, producing an optimism gap of 0.166 and a 26.8% decision flip rate. At $\gamma = 1.2$, development AUROC reaches 0.978 while replication AUROC falls to 0.753; the average optimism gap is 0.225 and 29.8% of decisions change.

Table 3. Illustrative Monte Carlo results across technical-confounding levels.

γ	Selected w	AUROC: Development $\rightarrow$ Replication	Optimism Gap; Decision Flips
0.0	0.010	0.864 $\rightarrow$ 0.864	0.000; 0.4%
0.4	0.411	0.893 $\rightarrow$ 0.837	0.057; 15.9%
0.8	0.827	0.944 $\rightarrow$ 0.778	0.166; 26.8%
1.2	0.996	0.978 $\rightarrow$ 0.753	0.225; 29.8%

Values are Monte Carlo means across 500 runs. *Fig. 2* displays the empirical 95% Monte Carlo intervals for AUROC and decision flip rates; full percentile intervals are reported in the text and simulation output. *Fig. 2* makes the managerial implication visible. Panel a shows that the internally optimized metric becomes more impressive as the technical confounder strengthens, while performance in the replication context moves in the opposite direction. Panel b shows that the downstream binary decision also becomes unstable. The result is intentionally stylized, but it demonstrates a general point: Model selection can reward a development-specific shortcut, so governance must ask how the decision behaves after the shortcut is removed or perturbed, not only how well the chosen pipeline fits its internal evaluation environment.

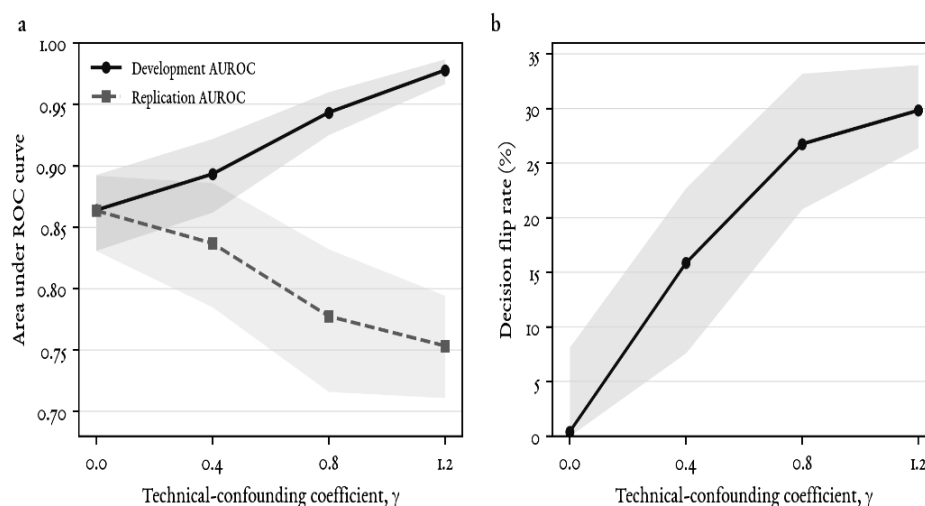


Fig. 2. Illustrative Monte Carlo stress test of internal performance and decision reproducibility.
Shaded regions are empirical 95% Monte Carlo intervals across 500 runs.

The simulation also provides a useful counterargument to indiscriminate skepticism about multi-omics modeling. At $\gamma = 0$, internal and replication performance are nearly identical, showing that analytical complexity is not itself the problem. The risk emerges when flexible modeling is allowed to exploit technical structure that is not stable under the intended use context. The appropriate response is therefore targeted stress testing rather than rejection of complex models.

4 | Risk-Based Governance for Multi-Omics Analytics

4.1 | Six Evidence Gates

The proposed governance architecture contains six evidence gates. A gate is not an additional committee by default; it is a minimum question that must be answered before the decision moves to a more consequential stage. The architecture is intentionally method-agnostic so it can sit above RNA sequencing, single-cell, proteomic, metabolomic, imaging, or integrated machine-learning workflows. *Fig. 3* summarizes the sequence and its feedback loop.

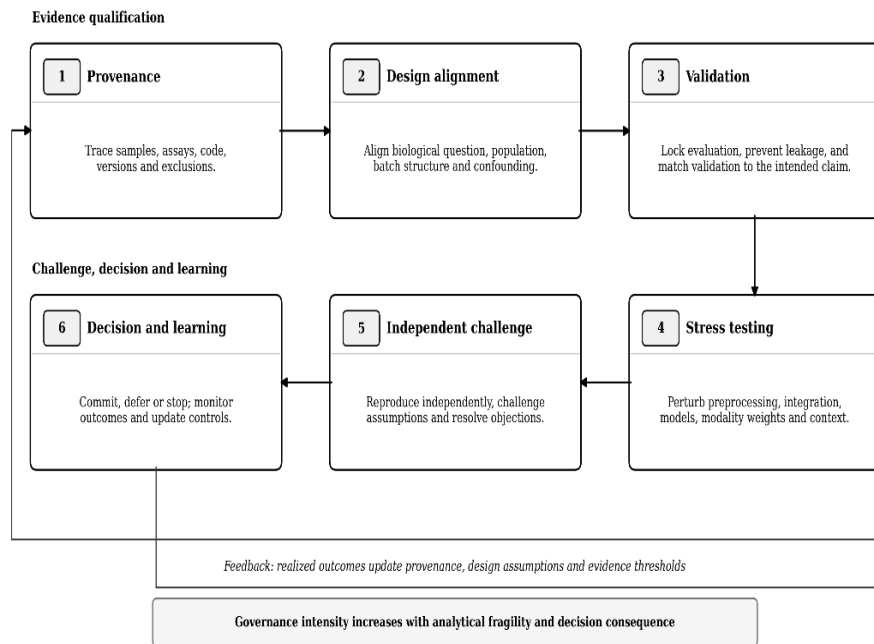


Fig. 3. Six evidence gates for risk-based governance of multi-omics analytics.

Table 4. Evidence gates, minimum evidence, and escalation triggers.

Gate	Minimum Evidence	Escalate When
1. Provenance	Versioned sample lineage, assay/batch metadata, exclusions, code environment, and transformations.	The analysis cannot be reconstructed or material lineage is missing.
2. Design alignment	Biological question, target population, assay rationale, blocking/randomization, and confounding review.	Biological and technical factors are inseparable or design cannot support the claim.
3. Validation	Locked evaluation, leakage audit, task metrics, uncertainty, and split logic matched to intended use.	Test reuse, cross-partition dependencies, or context-mismatched validation is detected.
4. Stress testing	Pre-specified perturbations to filtering, normalization, integration, models, modality weights, and context.	AFI, optimism gap, rank instability, or modality dependence exceeds tolerance.
5. Independent challenge	Independent reproduction, challenge memo, unresolved-failure log, and domain review.	Material objections remain unresolved or the independent result changes the decision.
6. Decision and learning	Commit/defer/stop rationale, owner, monitoring, realized outcome, and root-cause feedback.	Validity, performance, context, or consequence changes beyond tolerance.

4.2 | Risk-Adaptive Escalation

Governance intensity should increase with both analytical fragility and decision consequence. A useful operational starting point is an escalation priority E equal to consequence C multiplied by AFI, where both variables are scaled between 0 and 1. The score is not proposed as a universal regulatory threshold; it is a transparent prioritization device that makes the trade-off explicit.

$$E = C \times \text{AFI}, \quad 0 \leq C, \text{AFI} \leq 1. \quad (4)$$

Low-consequence, low-fragility analyses can proceed through standard pipeline controls. High-fragility but low-consequence work is suitable for rapid exploration because reversal is cheap and instability can be scientifically informative. Low-fragility but high-consequence decisions still warrant formal confirmation because the cost of a rare error is high. When consequence and fragility are jointly high, the default should move toward independent replication, additional data, deferral, or explicit senior escalation. This rule prevents two opposite failures: over-governing routine exploratory work and under-governing expensive or irreversible commitments.

4.3 | Decision Packages, Ownership, and Auditability

A project should not reach a consequential decision with only a slide deck of final model metrics. The minimum decision package should include the scientific question, cohort and assay map, provenance record, primary and stress-test analyses, decision-level stability measures, known limitations, unresolved challenges, and the explicit evidence that would cause the team to stop or reverse the decision. The package should be readable by both domain scientists and quantitative analysts, which is why decision-level measures such as AFI and optimism gap are valuable: They summarize robustness without hiding the underlying perturbation results.

Ownership should also be separated by function. The analysis team owns implementation and evidence production; the domain lead owns biological interpretation; an independent reviewer challenges hidden assumptions; and the project or portfolio owner accepts the residual risk of commitment. Combining all four roles in one person may be unavoidable in small laboratories, but the conflict should at least be visible. Independence is most important when the same team that optimized the model also benefits from a positive downstream decision.

5 | Managerial Implications and Research Agenda

5.1 | Operational Implications for Bioscience Organizations

The framework changes the target of analytics management. Instead of asking only which model is best, project leaders should ask which decision is sufficiently reproducible for the next resource commitment. This matters because the same analytical result can justify different governance responses at different stages. An exploratory pathway hypothesis may tolerate a high AFI if the next action is an inexpensive confirmatory assay; the same AFI should be unacceptable if it drives a costly validation program or a clinical-facing decision.

A second implication concerns portfolio efficiency. Reproducibility controls are often treated as costs added after analysis. In practice, provenance, blocked validation, perturbation testing, and independent challenge can reduce downstream rework by identifying fragile conclusions before they consume scarce experimental resources. The benefit is therefore not maximal certainty; it is earlier detection of assumptions that are cheap to fix now but expensive to discover later.

A third implication is organizational learning. A reproducibility incident should update the organization, not only the individual project. If a batch-confounding failure is diagnosed, future projects using the same

platform should inherit a design rule, a diagnostic test, or an automated warning. This converts post-decision monitoring from compliance into capability accumulation, consistent with organizational-learning theory [41].

5.2 | Testable Propositions

Proposition 1. Greater provenance completeness will be associated with shorter time-to-reproduce and shorter time-to-diagnose when conflicting multi-omics results arise.

Proposition 2. Projects that evaluate AFI across a pre-specified perturbation set will experience less downstream analytical rework than projects governed primarily by a single best-performing pipeline.

Proposition 3. The relationship between internal performance and downstream decision quality will weaken as batch, site, modality, and cohort heterogeneity increase unless validation explicitly represents those sources of variation.

Proposition 4. Independent analytical challenge will reduce optimistic commitment errors more strongly for high-complexity and high-consequence multi-omics projects than for low-complexity exploratory analyses.

Proposition 5. Risk-adaptive oversight will produce a better speed-quality trade-off than uniform review because it concentrates scarce expert attention on decisions with jointly high fragility and consequence.

Proposition 6. Organizations that encode diagnosed reproducibility failures into reusable design rules, pipeline checks, and training will show faster declines in recurring failure modes than organizations that treat incidents as project-specific exceptions.

5.3 | Maturity model and empirical designs

Implementation can be staged. Level 1, Reconstructable, requires traceable data, code, samples, versions, and exclusions. Level 2, Validated, protects evaluation from leakage and aligns validation with the intended claim. Level 3, Stress-tested, reports AFI, optimism gaps, modality dependence, and relevant context shifts. Level 4, Learning, links realized outcomes and incidents back to reusable organizational rules. The levels are cumulative: advanced model monitoring cannot compensate for missing provenance or a confounded study design.

The framework is empirically testable. A multicenter observational study could sample completed multi-omics projects and code provenance completeness, validation design, perturbation coverage, AFI, independent challenge, rework hours, replication outcome, and downstream assay success. Hierarchical models could estimate whether governance maturity predicts decision stability after controlling for sample size, modality count, research stage, and project complexity. A stronger prospective design could compare conventional review with risk-adaptive review on decision latency, expert-review hours, rework, and external validation loss. Computational benchmark studies can also report decision-level stability alongside predictive accuracy, especially when multiple pipelines are statistically close.

Boundary conditions matter. Some exploratory research has no stable decision threshold, so AFI may be premature. Very small samples can make perturbation results noisy. Platform-specific quality controls remain necessary and should not be replaced by a generic framework. Most importantly, instability can reveal real biological heterogeneity. A robust governance process must therefore ask whether the change is attributable to plausible biology, technical structure, or analyst choice rather than automatically treating all variation as error.

6 | Conclusion

Multi-omics analytics expands the biological questions that can be addressed, but it also expands the number of choices between raw measurement and scientific action. The resulting risk is not captured by accuracy alone. A model can look stronger internally because it has learned technical structure that disappears in a replication context, and a workflow can be perfectly rerunnable while the decision remains sensitive to

defensible alternative analyses. The paper therefore reframes reproducibility as both a decision-management problem and a computational one.

The proposed approach has three practical elements: measure decision-level analytical fragility, stress-test performance under validation and context changes that match the intended claim, and scale governance intensity to the joint level of fragility and decision consequence. The illustrative simulation shows why this matters: stronger technical confounding raised mean development AUROC from 0.864 to 0.978 while replication AUROC fell to 0.753 and nearly 30% of binary decisions changed. These values are not empirical estimates of real-world failure rates, but they demonstrate the mechanism that the governance framework is designed to detect.

The main limitation is that the quantitative results come from a deliberately simplified synthetic experiment and the governance thresholds have not yet been validated across real omics platforms, laboratories, or decision types. The literature synthesis is structured but not a systematic review, and AFI depends on the quality of the perturbation set chosen by analysts. Future work should validate the framework using multi-center bioscience projects, develop modality-specific perturbation libraries, study how AFI relates to replication and downstream experimental success, and test whether risk-adaptive oversight actually reduces total project cost and rework without suppressing productive exploration. A particularly important extension is longitudinal: research organizations should be studied as they reuse datasets, pipelines, and models over time to determine whether post-decision learning produces measurable gains in reproducibility and scientific efficiency.

Acknowledgments

None.

Author Contribution

Jamareh, N, and Jamareh, N, should review and confirm their individual CRediT roles before submission. Both authors must make substantive intellectual contributions, verify the analysis and references, revise the manuscript as needed, and approve the final submitted version.

Funding

Funding information was not provided during manuscript preparation and must be confirmed by the authors before submission.

Data Availability

No empirical human, animal, or patient dataset was analyzed. The quantitative results are generated from the synthetic Monte Carlo design described in Section 3. The authors should archive the final simulation code and analysis environment with the submitted manuscript or in a persistent repository.

Conflicts of Interest

The authors should verify and provide the journal-required conflict-of-interest statement before submission.

Generative Artificial Intelligence Disclosure

A large language model was used during manuscript preparation for literature organization, drafting assistance, code generation and debugging for the illustrative Monte Carlo simulation, figure drafting, and language editing. The authors are responsible for independently verifying all references, rerunning and validating the simulation, checking every numerical result and claim, making substantive revisions, and approving the final manuscript. The large language model is not an author.

References

- [1] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, 18, 83. <https://doi.org/10.1186/s13059-017-1215-1>
- [2] Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nature reviews genetics*, 19, 299-310. <https://doi.org/10.1038/nrg.2018.4>
- [3] Liu, F., Beck, S., Yang, L., Luo, H., & Zhang, K. (2026). Advancing AI for multi-omics and clinical data integration in basic and translational cancer research. *Nature reviews cancer*, 26, 497-512. <https://doi.org/10.1038/s41568-026-00922-2>
- [4] Cui, H., Tejada-Lapueta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F. J., & Wang, B. (2025). Towards multimodal foundation models in molecular cell biology. *Nature*, 640, 623-633. <https://doi.org/10.1038/s41586-025-08710-y>
- [5] Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews genetics*, 16, 85-97. <https://doi.org/10.1038/nrg3868>
- [6] Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and structural biotechnology journal*, 19, 3735-3746. <https://doi.org/10.1016/j.csbj.2021.06.030>
- [7] Acharya, D., & Mukhopadhyay, A. (2024). A comprehensive review of machine learning techniques for multi-omics data integration: Challenges and applications in precision oncology. *Briefings in functional genomics*, 23(5), 549-560. <https://doi.org/10.1093/bfpg/elae013>
- [8] Brooks, T. G., Lahens, N. F., Mrčela, A., & Grant, G. R. (2024). Challenges and best practices in omics benchmarking. *Nature reviews genetics*, 25(5), 326-339. <https://doi.org/10.1038/s41576-023-00679-6>
- [9] Hu, Y., Wan, S., Luo, Y., Li, Y., Wu, T., Deng, W., ... & Qu, K. (2024). Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature methods*, 21(11), 2182-2194. <https://doi.org/10.1038/s41592-024-02429-w>
- [10] Fu, S., Wang, S., Si, D., Li, G., Gao, Y., & Liu, Q. (2025). Benchmarking single-cell multi-modal data integrations. *Nature methods*, 22, 2437-2448. <https://doi.org/10.1038/s41592-025-02737-9>
- [11] Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for omics feature selection and multiple data integration. *PLOS computational biology*, 13(11), e1005752. <https://doi.org/10.1371/journal.pcbi.1005752>
- [12] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-omics factor analysis - a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6), e8124. <https://doi.org/10.15252/msb.20178124>
- [13] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21, 111. <https://doi.org/10.1186/s13059-020-02015-1>
- [14] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1), 13. <https://doi.org/10.1186/s13059-016-0881-8>
- [15] Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, 35(6), 498-507. <https://doi.org/10.1016/j.tibtech.2017.02.012>
- [16] Tung, P. Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., & Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7, 39921. <https://doi.org/10.1038/srep39921>
- [17] Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., ... & Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature reviews genetics*, 24(8), 550-572. <https://doi.org/10.1038/s41576-023-00586-w>
- [18] Bernett, J., Blumenthal, D. B., Grimm, D. G., Haselbeck, F., Joeres, R., Kalinina, O. V., & List, M. (2024). Guiding questions to avoid data leakage in biological machine learning applications. *Nature methods*, 21(8), 1444-1453. <https://doi.org/10.1038/s41592-024-02362-y>

- [19] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [20] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- [21] Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLOS computational biology*, 9(10), e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>
- [22] Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- [23] Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531-533. <https://doi.org/10.1038/483531a>
- [24] Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLOS biology*, 13(6), e1002165. <https://doi.org/10.1371/journal.pbio.1002165>
- [25] Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505, 612-613. <https://doi.org/10.1038/505612a>
- [26] McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., & Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. *Science translational medicine*, 13(586), eabb1655. <https://doi.org/10.1126/scitranslmed.abb1655>
- [27] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Aix-Covnet, Rudd, J. H. F., Sala, E., & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature machine intelligence*, 3, 199-217. <https://doi.org/10.1038/s42256-021-00307-0>
- [28] Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>
- [29] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Geburu, T. (2019). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). 2019 Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- [30] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2503-2511. <https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf>
- [31] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The lancet digital health*, 2(9), e489-e492. [https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2)
- [32] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England journal of medicine*, 385(3), 283-286. <https://doi.org/10.1056/NEJMc2104626>
- [33] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaneys-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature medicine*, 25, 1337-1340. <https://doi.org/10.1038/s41591-019-0548-6>
- [34] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17, 195. <https://doi.org/10.1186/s12916-019-1426-2>
- [35] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., & Reitsma, J. B. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [36] Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models

- using regression or artificial intelligence methods. *BMJ*, 388, e082505. <https://doi.org/10.1136/bmj-2024-082505>
- [37] Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388, e081554. <https://doi.org/10.1136/bmj-2024-081554>
- [38] World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization. https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content?utm_medium=email&utm_source=transaction
- [39] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
- [40] March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization science*, 2(1), 71-87. <https://doi.org/10.1287/orsc.2.1.71>
- [41] Argote, L., & Miron-Spektor, E. (2011). Organizational learning: From experience to knowledge. *Organization science*, 22(5), 1123-1137. <https://doi.org/10.1287/orsc.1100.0621>
- [42] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, 45(3), 1433-1450. <https://doi.org/10.25300/MISQ/2021/16274>
- [43] Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California management review*, 61(4), 66-83. <https://doi.org/10.1177/0008125619862257>
- [44] Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and organization*, 28(1), 62-70. <https://doi.org/10.1016/j.infoandorg.2018.02.005>