Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(1):90-100
https://doi.org/10.51847/C1XMwpDWx2

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

QSAR without False Confidence: A Model-Evidence Dossier for Applicability Domains, Calibration, Dataset Shift, and Decision-Relevant Uncertainty

João Silva[1]*, Pedro Costa[1], Ana Beatriz[2]

[1]Department of QSAR Modeling and Applicability Domains, Faculty of Pharmacy, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
[2]Department of Model Calibration and Uncertainty Quantification, Faculty of Pharmaceutical Sciences, University of Campinas, Campinas, Brazil.

***Email:** joao.silva@ufrj.br*

ABSTRACT

Quantitative structure activity relationship models support molecular prioritization, property estimation, virtual screening, and evidence preparation, yet apparently strong performance can create false confidence when detached from data provenance, endpoint meaning, chemical space coverage, applicability domains, calibration, dataset shift, and decision consequences. This article proposes the QSAR Model-Evidence Dossier, or Q-MED, as an original nonempirical architecture for organizing those dimensions around a declared intended use. Q-MED links an intended use contract, data and endpoint ledger, representation and modeling record, chemical space and applicability profile, calibration and uncertainty record, shift and challenge record, version and requalification log, and human accountability record through claim evidence traceability. A prediction is not decision eligible merely because an aggregate metric is favorable. Eligibility is conditioned on endpoint compatibility, relevant chemical support, evaluated uncertainty behavior, absence of unresolved shift, an identified model version, and accountable authorization. External validation and challenge sets test bounded claims across novel scaffolds, activity cliffs, sparse regions, altered assays, and other distributional changes. Monitoring, change control, and drift triggered requalification extend evidence assessment across the lifecycle. Defined human roles separate development, validation, scientific interpretation, decision authorization, monitoring, and retirement. Q-MED is a proposed conceptual synthesis, not a validated quality system, clinical recommendation, regulatory pathway, or deployment framework. Its value requires prospective evaluation of traceability, calibration, failure detection, selective prediction, external generalization, reviewer consistency, governance burden, and decision effects. It cannot compensate for invalid endpoints, biased data, inappropriate representations, unsupported causal interpretation, or inadequate experimental confirmation.

Key words: *QSAR, Pharmacoinformatics, Chemical space, External validation, Applicability domain, Calibration*

Received: 02 December 2024; **Revised:** 28 February 2025; **Accepted:** 01 March 2025

INTRODUCTION

Machine learning now contributes to target identification, molecular property prediction, virtual screening, lead prioritization, toxicity estimation, and other stages of pharmaceutical discovery and development. Within pharmacoinformatics, however, a model cannot be understood independently of the chemical and biological data on which it was trained, the molecular representation it receives, the prediction task it optimizes, and the evaluation design through which its performance is reported. Machine learning in this setting is therefore not merely an

algorithmic object but a coupled system of data, representation, model, validation procedure, and intended pharmaceutical use [1-3].

The widespread use of summary metrics can obscure this systems character. A high coefficient of determination, low mean error, favorable area under a receiver-operating-characteristic curve, or strong benchmark rank may be technically correct for the evaluated dataset while remaining uninformative about performance on new scaffolds, sparsely represented chemotypes, changed assays, shifted populations of compounds, or decisions with asymmetric consequences. Reported predictive success is particularly vulnerable to overinterpretation when it is treated as evidence of pharmaceutical usefulness without establishing whether the evaluation resembles the conditions under which the model will actually be used [2].

This article addresses that problem through the concept of false confidence: confidence in a QSAR output that exceeds the evidence available for the particular claim, compound, endpoint, model version, and decision context. False confidence does not require fraud, methodological negligence, or an intrinsically defective model. It can arise when a valid internal metric is extended beyond its evidentiary scope, when an applicability-domain designation is treated as a guarantee, when an uncertainty score is assumed to be calibrated, when a chemically similar compound is presumed to be predictively equivalent, or when an associational prediction is interpreted as evidence of intervention consequences. The central problem is consequently one of evidence alignment rather than performance measurement alone.

The objective is to develop an original, non-empirical evidence-governance architecture termed the QSAR Model-Evidence Dossier, or Q-MED. The proposed architecture connects intended use, data provenance, endpoint definition, molecular representation, chemical-space coverage, applicability-domain characterization, calibration, predictive uncertainty, distribution shift, external validation, version control, and human responsibility within a single claim-centered evidence record. This is an Original Evidence-Governance Architecture Article rather than an empirical validation study, regulatory guideline, software-validation report, or deployment manual. Q-MED is presented as an original conceptual synthesis requiring computational, experimental, organizational, and decision-process evaluation before its utility or readiness can be established.

Why performance metrics do not constitute sufficient qsar evidence

Apparently strong molecular model performance may coexist with benchmark memorization, data leakage, or sensitivity to dataset, split, representation, and optimization choices [4-6]. These mechanisms threaten different evidentiary claims. Memorization rewards familiarity, leakage allows unavailable information to influence evaluation, and pipeline choices determine which relationships are tested and which errors are hidden by aggregation.

The distinction between interpolation and prospective generalization is central. Random partitioning may place close analogues or members of one series on both sides of the split. The resulting estimate can be useful for interpolation, yet it may not represent transfer to novel scaffolds or sparsely sampled regions. Random splits are not inherently invalid; their claim must be bounded [4].

Performance also depends on molecular standardization, label aggregation, descriptor or graph construction, hyperparameter selection, and test design. Comparative evidence shows that these choices interact [6]. Consequently, a score belongs to a specific data and modeling pipeline. It should not be treated as a portable property of a model family or software implementation.

Q-MED replaces metric centered evidence with claim centered evidence. Each claim identifies the endpoint, chemical population, data conditions, model version, and decision supported. Aggregate metrics remain necessary when relevant, but they cannot substitute for local support, calibration, external validity, shift resistance, mechanistic interpretation, or decision benefit. The critique concerns inference from metrics, not the usefulness of metrics themselves.

Intended use and decision context

The adequacy of a QSAR output depends on what it is expected to do. Uncertainty becomes decision relevant only when linked to an action, error consequence, or rule for withholding a prediction. The same output may support exploratory prioritization but remain insufficient for exclusion, escalation, or another consequential decision [7].

Q-MED begins with an intended use contract specifying the endpoint, compound population, scientific user, pharmaceutical stage, permissible use, prohibited use, downstream confirmation, and consequences of error.

Assay format, biological system, units, curation, and sampling influence what labels mean and therefore constrain admissible use [8].

The contract is paired with minimum information that reconstructs relationships among data, model, evaluation, and claim. Reporting principles support documentation of purpose, data source, preprocessing, model construction, evaluation, limitations, and intended interpretation [9]. In Q-MED, these become versioned evidence fields rather than publication statements alone.

The architecture distinguishes exploratory mapping, ranking, quantitative estimation, exclusion, and higher consequence support. These are proposed use classes, not validated risk levels. Evidence should be proportional to reversibility, experimental confirmation, error asymmetry, and the cost of an incorrect decision. A model may be eligible for one use while remaining ineligible for another.

Architecture of the proposed model-evidence dossier

A defensible QSAR record must preserve reproducibility information, the staged model building workflow, and chemistry specific data and evaluation practices rather than only a final model object [10-12]. Q-MED extends this documentation into claim specific governance, where each evidentiary object is connected to a defined use, model version, limitation, and accountable owner.

Figure 1 presents the proposed Q-MED architecture, in which data, model, applicability, uncertainty, validation, lifecycle, and accountability evidence converge on an intended use specific decision gate.

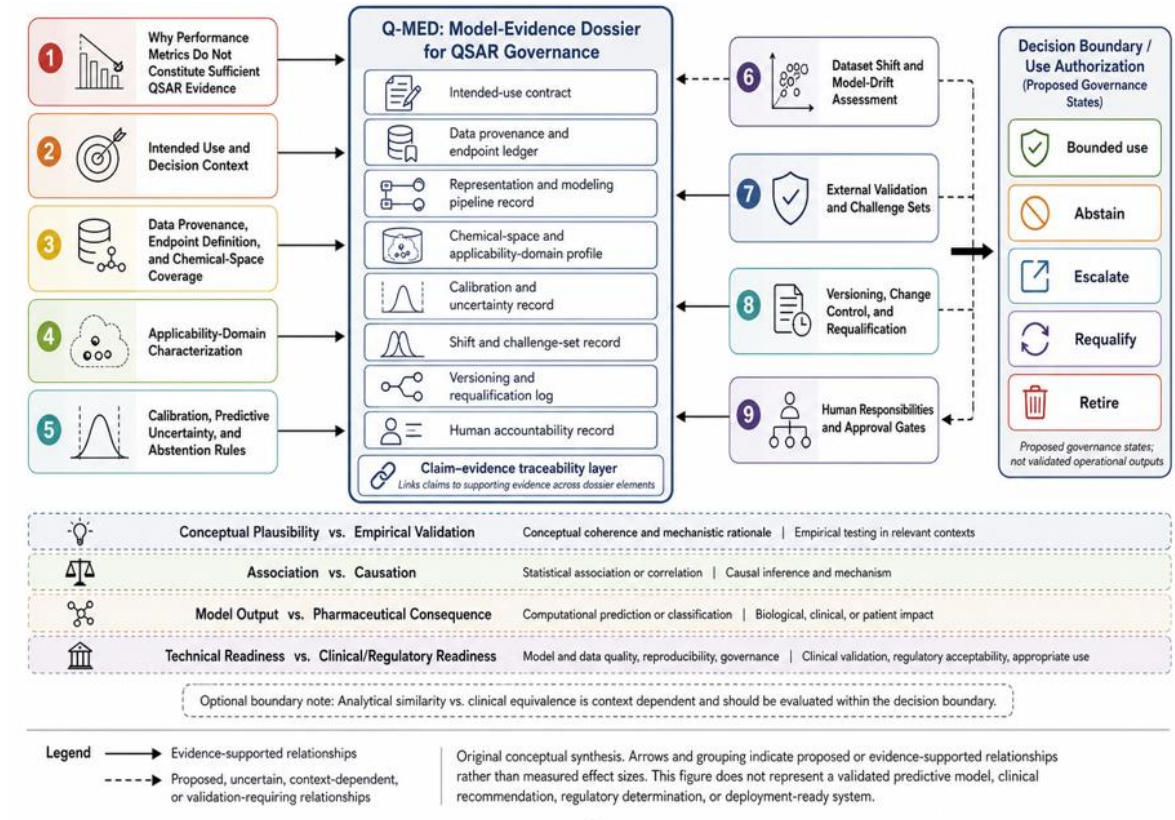

Figure 1. Model-Evidence Dossier Architecture for QSAR Governance.

The figure is an original conceptual synthesis developed for "QSAR without False Confidence: A Model-Evidence Dossier for Applicability Domains, Calibration, Dataset Shift, and Decision-Relevant Uncertainty." Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, clinical recommendation, regulatory determination, or deployment-ready system.

The proposed dossier contains eight modules: intended use; provenance and endpoint definition; representation and pipeline; chemical space and applicability; calibration and uncertainty; shift and challenge evidence; versioning and requalification; and human accountability. A cross cutting traceability layer connects claims to evidence and unresolved limitations.

The modules are interdependent. Intended use determines which errors and tests matter. Provenance defines what labels and compounds represent. Representation shapes similarity and applicability. Calibration and external validation qualify confidence. Monitoring can invalidate earlier evidence. Human authorization integrates these records but cannot repair weak science.

Evidence is stored as structured objects, including dataset lineage, endpoint rules, split rationale, local support analyses, calibration results, challenge outcomes, monitoring signals, and approval decisions. Evidence may be present yet insufficient, outdated, inconsistent, or irrelevant to the requested use. Q-MED is designed to make those states visible.

A proposed decision gate permits bounded use only when intended use, endpoint context, applicability, uncertainty, shift status, version, and human approval are compatible. Otherwise, the response is abstention, escalation, requalification, or retirement. These states organize decisions; they are not validated thresholds or guarantees of correct outcomes.

Data provenance, endpoint definition, and chemical-space coverage

Data provenance determines what a QSAR dataset can legitimately support. Source databases, assay programs, compound-selection strategies, molecular standardization, duplicate handling, and training–test construction shape both the patterns available to the model and the apparent difficulty of the evaluation task. Dataset assembly and cross-docking choices can materially alter what is learned and what reported validation represents [13].

Endpoint definition must preserve the scientific meaning of each label. A numerical value can conceal differences in assay format, biological system, units, censoring, replicate aggregation, detection limits, and experimental protocol. Chemical datasets may also contain source-related or design-related biases that enable models to exploit unintended signals rather than transferable structure–activity relationships [14]. Q-MED therefore requires endpoint definitions to remain linked to their experimental and curation context.

Chemical-space coverage is distinct from dataset size. A large dataset may remain concentrated around a limited set of scaffolds, property ranges, or medicinal-chemistry programs, leaving other regions sparsely supported. Coverage bias can produce uneven learning conditions and misleading aggregate evaluation when densely represented regions dominate the reported result [15]. The dossier should therefore record scaffold distribution, local density, property coverage, and known underrepresented regions.

These records do not prove that a prediction is reliable. They establish the evidentiary conditions against which applicability, uncertainty, and external validity must subsequently be judged.

Table 1 organizes the principal constructs, evidence objects, relationships, intended uses, uncertainties, and interpretation boundaries in Q-MED.

Table 1. Definitions, components, relationships, and intended uses in QSAR without False Confidence.

Component	Problem addressed	Inputs	Proposed relationship	Expected contribution	Evidence required	Failure risk	Boundary
Intended use contract	Claims lack decision context	Endpoint, user, population, consequences	Proposed: binds claim to permitted use	Decision specific evidence	Explicit use and limitations [9]	Informal scope expansion	A stated use is not validated use
Provenance ledger	Data lineage is hidden	Sources, curation, splits	Proposed: links claims to exact data	Traceability	Reconstructable assembly [13]	Bias remains undocumented	Provenance does not correct data
Endpoint record	Labels are treated as equivalent	Assay, units, censoring	Proposed: preserves scientific meaning	Limits unsupported transfer	Assay and curation metadata [14]	Heterogeneity persists	Precision does not establish relevance
Chemical coverage	Dataset size masks gaps	Scaffolds, density, properties	Proposed: maps supported regions	Visible evidence gaps	Coverage analysis [15]	Representation dependence	Coverage does not guarantee accuracy
Applicability profile	One domain rule creates certainty	Similarity, density, leverage	Proposed: preserves multiple signals	Compound specific support	Indicator error relationship [16]	Signals conflict	In domain is not error free

Calibration record	Confidence is unevaluated	Probabilities, intervals, outcomes	Proposed: links confidence to error	Supports abstention	Calibration and coverage tests [17]	Shift breaks calibration	Confidence is not decision benefit
Shift record	Validation becomes stale	Inputs, assays, outcomes, context	Proposed: links signals to claims	Lifecycle investigation	Defined OOD assessment [18]	False alarms or missed drift	A signal is not proof of failure
Version and roles	Claims outlive model changes	Data, code, approvals, owners	Proposed: binds use to version	Accountable change control	Maintenance record [19]	Symbolic governance	Documentation is not qualification

Applicability-domain characterization

An applicability domain describes conditions under which a model has relevant empirical support, but no single definition captures every form of support. Similarity, leverage, density, descriptor range, and model based indicators answer different questions. Comparative evidence does not support one universally superior applicability measure [16].

Ensemble disagreement may indicate local instability or sensitivity to training variation. Its usefulness depends on the task, ensemble construction, and empirical relationship between disagreement and error. Large scale evaluation shows that ensemble variation should be tested rather than assumed to provide a complete applicability criterion [20].

Uncertainty in property estimation may also reflect sparse chemical or property support. QSAR oriented methods can identify predictions produced under weak local conditions [21]. However, closeness in one representation does not establish mechanistic similarity, assay equivalence, or equal pharmaceutical consequences.

Q-MED proposes a plural applicability profile recording each indicator, representation, validation evidence, model version, compound result, and disagreement. Discordant signals are preserved rather than collapsed into an unsupported score. The purpose is to reveal strong, weak, conflicting, or absent support, not to certify accuracy.

Calibration, predictive uncertainty, and abstention rules

Calibration asks whether stated confidence corresponds to observed outcomes under specified conditions. It differs from average accuracy: a model may rank compounds effectively while assigning poorly calibrated probabilities or intervals. Molecular uncertainty methods differ in calibration, error association, and behavior under domain change [17].

Producing a variance, interval, or ensemble spread is therefore insufficient. Scalable methods differ in identifying large errors, maintaining coverage, and remaining informative outside familiar regions [22]. Q-MED records the method, calibration data, evaluation criterion, version, recalibration history, and known failure conditions.

Local conformal prediction illustrates how intervals or prediction sets can be linked to explicit coverage objectives and local applicability conditions [23]. Such guarantees remain conditional on assumptions and future data exchangeability. They do not establish mechanism, clinical benefit, or reliability under untested shifts.

Q-MED attaches abstention rules to intended use. A prediction may be withheld when calibration evidence is inadequate, applicability indicators conflict, intervals are unsuitable for the decision, or the query lies outside supported conditions. No universal threshold is proposed. Each rule requires prospective evaluation of avoided errors, missed opportunities, experimental burden, and decision consequences.

Dataset shift and model-drift assessment

Activity cliffs show that structurally similar compounds can have sharply different activities, producing local errors hidden by aggregate metrics. These discontinuities challenge smooth similarity assumptions and reveal why near neighbor confidence may be misplaced [24].

Out of distribution status is not one property. A compound may be novel by scaffold, representation, target context, assay source, property range, or time. Real world molecular OOD evaluation shows that these specifications create different generalization questions and must be declared explicitly [18].

Dataset shift should be separated from model drift. Dataset shift concerns changed data generating conditions or observed distributions; model drift concerns deterioration in the relationship among inputs, predictions, endpoints, and use. Causal structure can help distinguish superficial changes from changes threatening the predictive relationship [25].

Q-MED proposes a shift register covering input chemistry, endpoint or assay, processing and curation, observed performance, and evidence changes caused by new knowledge. Each signal is linked to affected claims. Detection begins investigation; it does not by itself establish failure or prescribe a response.

Monitoring compares current inputs, outcomes, calibration, applicability, and use with qualified conditions. Signals include scaffold changes, altered density, unexpected local errors, assay modification, or prediction observation divergence. The architecture requires verification, impact assessment, and documented uncertainty before continuation, restriction, revalidation, or retirement.

Table 2 translates the architecture into testable processes, validation criteria, failure modes, responsibilities, and readiness boundaries.

Table 2. Testable propositions, evidence requirements, and validation criteria for QSAR without False Confidence.

Stage	Core function	Information flow	Interaction	Validation criterion	Failure mode	Responsibility	Readiness boundary
Intended use	Bound the claim	Decision context to contract	Sets later evidence needs	Reviewer agreement	Vague scope	Domain and decision owners	No consequential use without explicit scope
Data qualification	Establish learned evidence	Measurements to curated labels	Feeds coverage and validation	Reconstructable lineage	Bias or leakage	Data owner and experimental reviewer	Qualified data are not valid endpoints
Model development	Create reproducible prediction	Structures to representation and model	Shapes applicability and uncertainty	Independent reconstruction	Missing chemistry	Developer and validator	Reproducibility is not generalization
Applicability and calibration	Assess local support	Query and outputs to support evidence	Drives abstention	Error association and coverage	False reassurance	Validator and domain reviewer	No universal threshold
External challenge	Attempt claim falsification	Independent cases to failure analysis	Tests bias, shift, and calibration	Prespecified independent evaluation	Challenge repeats training bias	Experimental owner and validator	One pass is not universal transportability
Monitoring and requalification	Maintain bounded validity	Use evidence to impact assessment	Updates affected modules and approvals	Detection and justified response	Missed or excessive triggers	Monitoring owner and decision owner	Proposed process requiring validation

External validation and challenge sets

External validation tests whether claims survive conditions not used for construction or tuning. Generalizability depends on relationships among compounds, labels, sources, and distributions and cannot be inferred from one internal split [26]. The independence, relevance, and difficulty of every external dataset should therefore be documented.

Figure 2 presents the lifecycle pathway from model development through bounded use, monitoring, shift investigation, and drift triggered requalification.

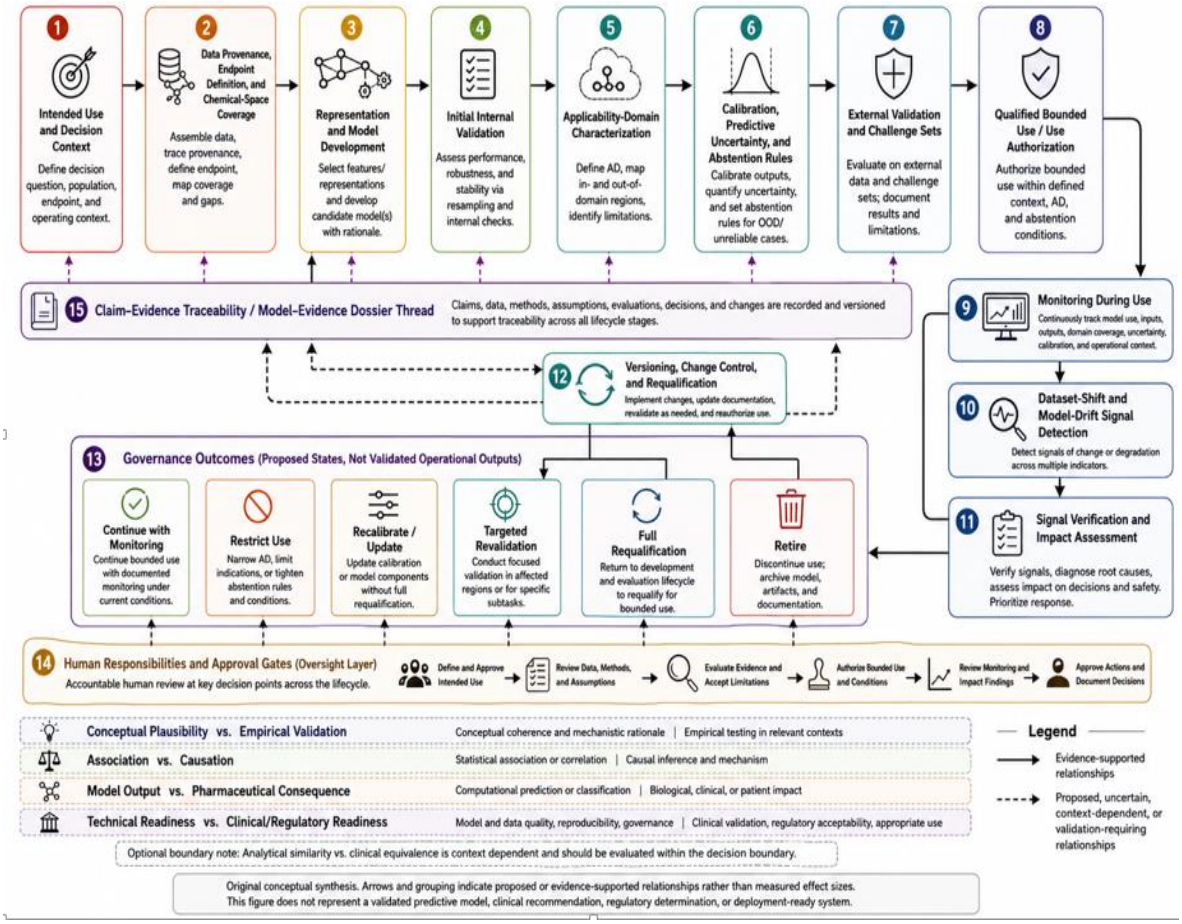

Figure 2. Lifecycle Pathway from Model Development to Drift-Triggered Requalification

This original conceptual synthesis connects intended use, data assembly, model development, internal and external validation, bounded use, monitoring, shift investigation, impact assessment, requalification, and retirement. Arrows represent logical or proposed relationships rather than measured effects or approved procedures. No universal trigger, interval, threshold, or readiness claim is implied.

External data are not automatically unbiased. Shared series, common assay origins, duplicated measurements, or related curation can reproduce development data structure. Binding affinity evidence shows that controlling dataset bias can change apparent generalization [27]. Q-MED distinguishes temporal, scaffold, assay, source, and institutional independence.

Challenge sets intentionally concentrate difficult conditions such as novel scaffolds, activity cliffs, sparse regions, conflicting applicability indicators, altered assays, uncertain labels, or known confounders. Their purpose is attempted falsification, not representative performance estimation. A model may pass a broad external set while failing a scientifically important challenge.

External success remains conditional evidence. It supports tested claims but cannot certify universal performance across untested chemical, biological, temporal, or operational settings. Experimental confirmation is required when the claim depends on unseen molecular behavior rather than computational consistency alone, and failure analysis should remain visible beside aggregate results.

Versioning, change control, and requalification

Qualification is bounded to a combination of data, endpoint rules, representation, code, parameters, software dependencies, and intended use. Predictive systems require continuing ownership, monitoring, and maintenance rather than one time validation [19]. Every claim in Q-MED is therefore attached to a specific version.

Change control distinguishes administrative edits from scientifically material changes. New data, revised labels, altered standardization, representation changes, code modifications, retraining, recalibration, or expanded use can affect several evidence modules. Validation and monitoring practice supports documented reassessment rather than automatic transfer of prior evidence [28].

Q-MED proposes impact based requalification. Each change is mapped to affected claims, modules, tests, and decision gates. A narrow change may justify targeted reassessment, whereas interacting changes may require renewed external validation, applicability analysis, calibration, and challenge testing. No universal definition of materiality is asserted.

Requalification may support continued use, restriction, recalibration, retraining, expanded validation, suspension, or retirement. These are proposed governance responses rather than approved procedures. The scope, rationale, unresolved uncertainties, new version, and accountable authorization must be recorded before renewed use. T**able 3** consolidates decision tradeoffs, triggers, evidence needs, uncertainty treatments, corrective actions, and prohibited conclusions.

Table 3. Failure modes, boundary conditions, governance needs, and implementation priorities for QSAR without False Confidence.

Scenario	Trigger	Competing objective	Evidence threshold	Uncertainty treatment	Corrective action	Expected implication	Cannot conclude
Metric gain versus generalization	Internal score improves	Optimization versus realism	Leakage resistant evaluation [4]	Treat unexplained gain cautiously	Audit splits and retest	Clarifies source of gain	Higher score means better decisions
Coverage versus consistency	New sources added	Breadth versus endpoint comparability	Provenance compatibility [15]	Record assay heterogeneity	Stratify or restrict claims	May broaden evidence	More data ensure validity
Single domain rule versus plurality	Indicators disagree	Simplicity versus completeness	Empirical indicator evaluation [20]	Preserve disagreement	Escalate or abstain	Makes conflict visible	Agreement guarantees accuracy
Precision versus calibration	Intervals are narrow	Apparent certainty versus coverage	Calibration evidence [22]	Treat unvalidated precision as uncertain	Recalibrate or withhold use	Reduces unsupported certainty	Narrow intervals are reliable
Continue versus investigate shift	Inputs or assays change	Continuity versus claim protection	Verified material impact [25]	Separate signal and impact uncertainty	Targeted testing or restriction	Prevents automatic continuation	Distribution change proves failure
Maintain versus requalify	Data, code, or use changes	Efficiency versus evidence continuity	Mapped material change [28]	Record interacting unknowns	Targeted or full reassessment	Keeps claims version bound	Prior evidence transfers automatically
Explanation versus fidelity	Attribution appears plausible	Interpretability versus faithfulness	Stability and fidelity tests [29]	Label weak explanations uncertain	Reject as approval evidence	Limits decorative interpretation	Explanation establishes mechanism
Prediction versus intervention	Association guides action	Usefulness versus causal overreach	Separate causal evidence [30]	Record unsupported assumptions	Restrict claim or test causally	Preserves decision boundary	Prediction establishes intervention benefit

Human responsibilities and approval gates

Evidence governance requires identifiable responsibility across data creation, model development, validation, decision use, monitoring, and failure response. Responsible machine learning principles place obligations across the lifecycle rather than solely on developers [31]. Q-MED makes those responsibilities explicit and traceable.

The proposed roles are data owner, model owner, independent validator, domain reviewer, decision owner, and monitoring owner. One person may hold several roles in a small organization, but each judgment remains attributed. Independence should increase when decisions are expensive, irreversible, or exclusionary.

Interpretability requires a separate approval question. QSAR explanations may appear chemically plausible without being faithful, stable, or actionable. Benchmarking shows that explanation quality requires explicit evaluation rather than visual acceptance [29]. Explanations are therefore evidence objects with limitations, not automatic mechanistic justification.

Approval gates integrate evidence but do not replace it. The decision owner may authorize bounded use, impose restrictions, request experiments, trigger requalification, or retire a claim. Human participation can introduce automation bias, inconsistency, and responsibility diffusion; approval quality must itself be assessed through traceability, disagreement, escalation, and detection of unsupported use.

Limitations and Implementation Priorities

Q-MED does not convert association into causal knowledge. A model predicting activity, toxicity, solubility, or exposure related properties does not establish what would happen under an intervention or which action is optimal. Causal and counterfactual decisions require assumptions and evidence beyond predictive association [30].

The architecture cannot repair invalid endpoints, hidden bias, missing chemistry, or inadequate experiments by documenting them. Data quality, provenance, coverage, and availability remain fundamental constraints on molecular machine learning [32]. A complete dossier may reveal uncertainty without resolving it.

The framework also creates organizational risks. Modules may overlap, evidence requirements may be costly, and reviewers may disagree about sufficiency. Formal documentation can become performative, encourage checklist compliance, or create another form of false confidence. Applicability, calibration, challenge, and monitoring evidence remain model and context dependent.

Implementation should begin with bounded comparative studies. Evaluation should test whether Q-MED improves traceability, detects failures missed by metric only review, supports appropriate abstention, recognizes material drift, improves reviewer consistency, and avoids disproportionate burden. Adoption should depend on demonstrated value and unintended effects, not conceptual completeness alone.

Prospective evaluation should examine whether reviewers can identify unsupported claim extension, distinguish missing evidence from negative evidence, preserve dissent, document override reasons, and recognize when additional experiments are more informative than further modeling. Comparative studies should include different organizational settings, endpoint classes, data volumes, model families, and decision consequences. They should report detected failures together with false alarms, delayed decisions, unnecessary experiments, reviewer fatigue, maintenance cost, and instances in which documentation increases confidence without improving judgment. Particular attention is needed for small teams that cannot separate every governance role, legacy models with incomplete provenance, continuously updated datasets, and applications where experimental confirmation is slow, expensive, or ethically constrained. These studies can clarify which dossier modules are essential, which can be simplified, and which require stronger independence, automation, or technical support.

Evaluation should also test whether abstention and escalation rules remain understandable to domain experts, whether version changes are recognized consistently, and whether challenge sets reveal failures that ordinary validation overlooks. Governance burden should be treated as an outcome, because a framework that is too complex to maintain may encourage superficial compliance or undocumented workarounds. The relevant comparator is not an idealized dossier against no governance, but Q-MED against the documentation, review, validation, and maintenance practices already used in the intended setting. Evidence of benefit should remain specific to the claims, workflows, users, compounds, endpoints, and consequences studied. Negative or mixed findings should guide revision rather than be hidden behind formal completeness. Further research should assess training needs, software interoperability, data access constraints, dispute resolution, auditability, role conflicts, and conditions under which retirement is preferable to repeated modification. Transparent reporting of unsuccessful pilots will be especially important for preventing the architecture itself from becoming a new source of false confidence. Implementation decisions should ultimately ask whether the dossier improves scientific reasoning, makes uncertainty visible, and supports proportionate action without displacing experimental judgment or creating administrative certainty unsupported by evidence.

Prospective designs should therefore predefine success and failure criteria, preserve all reviewer disagreements, and distinguish improvements in record completeness from improvements in decisions. Studies should track how often users consult the dossier, which modules influence action, where information becomes stale, and whether model owners respond appropriately to new evidence. They should also examine unintended incentives, including avoidance of difficult compounds, excessive abstention, strategic selection of favorable challenge sets, and reluctance to retire familiar models despite weakened support across diverse practical pharmaceutical research settings.

CONCLUSION

QSAR without false confidence requires evidence beyond aggregate predictive performance. Q-MED organizes intended use, provenance, endpoint meaning, representation, chemical coverage, applicability, calibration, uncertainty, shift, external validation, version control, and accountability around traceable claims. Its decision gate permits bounded use, abstention, escalation, requalification, or retirement. The architecture is an original conceptual synthesis, not a validated quality system, regulatory pathway, clinical recommendation, or deployment ready solution. Its value depends on prospective computational, experimental, organizational, and decision process evaluation. Even faithful implementation cannot substitute for valid endpoints, representative data, appropriate modeling, independent evidence, experimental confirmation, or causal reasoning when a claim exceeds prediction.

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, Cortés-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. Rodríguez-Pérez R, Miljković F, Bajorath J. Machine learning in chemoinformatics and medicinal chemistry. Annu Rev Biomed Data Sci. 2022;5:43-65. doi:10.1146/annurev-biodatasci-122120-124216
4. 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
5. 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
6. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
7. Yu J, Wang D, Zheng M. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? iScience. 2022;25(8):104814. doi:10.1016/j.isci.2022.104814
8. Bender A, Cortés-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
9. 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
10. 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
11. Kausar S, Falcão AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1):1. doi:10.1186/s13321-017-0256-5
12. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
13. Francoeur PG, Masuda T, Sunseri J, Jia A, Iovanisci RB, Snyder I, et al. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. J Chem Inf Model. 2020;60(9):4200-15. doi:10.1021/acs.jcim.0c00411
14. 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
15. Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S. Coverage bias in small molecule machine learning. Nat Commun. 2025;16(1):554. doi:10.1038/s41467-024-55462-w

16. Klingspohn W, Mathea M, ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. J Cheminform. 2017;9(1):44. doi:10.1186/s13321-017-0230-2
17. Dutschmann TM, Kinzel L, ter Laak A, Baumann K. Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. J Cheminform. 2023;15(1):49. doi:10.1186/s13321-023-00709-9
18. Brown TN, Sangion A, Arnot JA. Identifying uncertainty in physical-chemical property estimation with IFSQSAR. J Cheminform. 2024;16(1):65. doi:10.1186/s13321-024-00853-w
19. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
20. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
21. Oršolić D, Šmuc T. Dynamic applicability domain (dAD): Compound-target binding affinity estimates with local conformal prediction. Bioinformatics. 2023;39(8). doi:10.1093/bioinformatics/btad465
22. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
23. Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. J Chem Inf Model. 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
24. 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
25. Ektefaie Y, Shen A, Bykova D, Marin MG, Zitnik M, Farhat MR. Evaluating generalizability of artificial intelligence models for molecular datasets. Nat Mach Intell. 2024;6(12):1512-24. doi:10.1038/s42256-024-00931-6
26. Graber D, Stockinger P, Meyer F, Mishra S, Horn C, Buller R. Resolving data bias improves generalization in binding affinity prediction. Nat Mach Intell. 2025;7(10):1713-25. doi:10.1038/s42256-025-01124-5
27. Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools: A 360° approach to model maintenance. J Am Med Inform Assoc. 2024;31(5):1195-8. doi:10.1093/jamia/ocae036
28. Spies NC, Farnsworth CW, Wheeler S, McCudden CR. Validating, implementing, and monitoring machine learning solutions in the clinical laboratory safely and effectively. Clin Chem. 2024;70(11):1334-43. doi:10.1093/clinchem/hvae126
29. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
30. Matveieva M, Polishchuk P. Benchmarks for interpretation of QSAR models. J Cheminform. 2021;13(1):41. doi:10.1186/s13321-021-00519-x
31. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369-75. doi:10.1038/s42256-020-0197-y
32. Durant G, Boyles F, Birchall K, Deane CM. The future of machine learning for small-molecule drug discovery will be driven by data. Nat Comput Sci. 2024;4(10):735-43. doi:10.1038/s43588-024-00699-0