The Supervisory Evidence Gap

SUPERVISORY TECHNOLOGY | SHARED EVIDENCE | DECISION QUALITY | EVALUATION

ABSTRACT

More frequent regulatory reporting does not necessarily produce more timely or better-supported supervisory decisions. Supervisors must also identify how new evidence affects prior assessments. This paper develops a shared evidence record linking source information, access permissions, unresolved questions and supervisory decisions. A critical literature review situates the proposal within research on regulatory reporting, evidence provenance and the revision of decisions when their supporting assumptions change. An executable model specifies how changes in evidence or permissions identify decisions requiring review. It passes 184 checks and detects eight deliberately introduced faults; these results establish selected properties of the model, rather than effectiveness in regulatory practice. A prospective comparative study holds evidence availability, delivery times, access rights and analytical capabilities constant to assess the contribution of linking evidence to decisions. Its primary hypothesis concerns a reduction in time to justified initial and revised decisions, subject to predefined conditions for decision quality, workload and disclosure. The statistical analysis plan and synthetic simulations assess these conditions jointly and identify sensitivity to sample size and missing observations. The paper establishes a conceptual and testable basis for linking shared evidence to supervisory decisions. Effects on human performance and total regulatory burden require independent empirical evaluation.

Keywords: continuous supervision; supervisory technology; regulatory reporting; evidence integration; data governance; cross-organisational information sharing

1 Introduction

A supervisor may receive a report of a service failure without knowing whether it changes an earlier risk assessment. Two reports may both contain current figures but cover different parts of a banking group: one may report lending across the group, while another covers a single subsidiary. Comparing these figures without accounting for their different coverage may produce an erroneous assessment. A firm may also mark corrective work as complete while the evidence that the problem has been resolved remains disputed. These examples distinguish information acquisition from its interpretation in relation to a supervisory decision.

Financial supervision is a demanding setting for studying how organisations share and use evidence. Regulated firms produce much of the information, while supervisors judge its significance under their responsibilities. Evidence sharing can preserve differences in interpretation, access rights and decision-making authority between organisations.

The research question is:

> *When evidence, access rights and delivery times are held constant, does linking evidence to prior decisions reduce the time and effort required to make and revise justified supervisory assessments?*

A secondary question concerns the effect of earlier evidence delivery. Separating these questions allows the study to distinguish the contribution of delivery timing from that of evidence organisation.

Here, "continuous supervision" means reviewing assessments as relevant information changes, using an agreed update schedule. This definition does not imply unrestricted access to firms' records, uninterrupted data transmission or automatic certification of regulatory compliance. Regular reports, enquiries, inspections and professional judgement can remain part of supervision.

The paper makes three contributions:

1. A specification of a shared evidence record that distinguishes a decision's historical evidential basis from its continuing validity.
2. An executable model and failure tests that demonstrate selected properties within explicit implementation limits.
3. A prospective evaluation design that separates evidence organisation from delivery timing and assesses decision quality, workload and disclosure alongside response time.

Appendix A summarises the model and its assumptions. Appendix B connects failure cases to the available checks. Appendix C explains the main terms used in the manuscript and technical supplement.

2 Literature review and research contribution

2.1 Review scope and method

The selected review covers continuous supervision, supervisory technology, digital regulatory reporting, evidence integration and methods for recording why a decision was made. It also covers controls on who may use information and for what purpose, and how earlier conclusions are reviewed when their supporting assumptions change. The reference list identifies the versions used; the source register distinguishes verified publication records from full-text consultation. Nine cited papers available through arXiv address related methods; several concern data governance more broadly rather than supervision itself. The supplement records the sources consulted and access limits. Publication dates are restricted to material available by 1 September 2026; editorial verification and reproducibility checks were completed subsequently. This was a selected critical review, not an exhaustive systematic review.

The evidential contribution of each source depends on its design and status. Institutional accounts document existing or planned systems but do not, in themselves, establish causal effects. Prototype reports provide evidence of implementation and exploratory evaluation. Preprints may not have undergone peer review. Where only an abstract was accessible, the interpretation is restricted to its reported claims. Consistent with design-science research, this paper distinguishes model construction from evaluation of its practical utility [19].

2.2 Reporting and supervisory judgement

Research on supervisory technology, often called SupTech, distinguishes collecting data from analysing it and identifies practical implementation difficulties [3, 6, 9]. The Digital Regulatory Reporting Phase 2 Viability Assessment examines reporting designs and costs [13]. Its estimates draw on limited industry input, and the report does not state a binding regulatory position. Automated reporting and common reporting definitions are therefore established ideas.

Research on responsive risk-based regulation distinguishes finding a problem, responding to it, assessing the response and adapting the approach [5]. This supports evaluating the justification and revision of supervisory decisions, rather than treating alert volume as a measure of supervisory effectiveness. The present paper addresses a narrower question: how teams handle information about operational incidents, such as service failures, and the work undertaken to resolve them.

Embedded supervision studies monitoring based on information held in distributed ledgers and the conditions under which that information can be trusted [1]. DLT Compliance Reporting develops and evaluates a system for collecting and adding context to information with little delay [2]. Both examine ways to reduce the steps between firms' records and supervisory information. This paper does not require a distributed ledger. Its focus is the effect of evidence organisation on initial assessment and reassessment when information availability is held constant.

2.3 Existing supervisory systems

Project Ellipse is a prototype that brings regulatory data and analysis together [4]. The European Central Bank's 2024 supervisory report describes digital processes for receiving incident reports and monitoring corrective work, alongside plans for a unified supervisory interface [15]. Its 2025 report records continued work on Olympus and this interface [16]. Information integration is therefore an established capability rather than, by itself, an original contribution. These accounts also provide no basis for assuming that other regulators lack comparable capabilities.

The proposed contribution is a controlled evaluation of the value of explicit links between evidence and decisions. Teams would receive equally current information, with the same access rights and analytical tools, but use different ways of organising evidence and linking it to decisions. If a regulator's existing system already provides the proposed connections, it belongs to the class of systems being studied. The experiment cannot test the benefit of adding those connections where there is no meaningful difference to compare.

Implementation also depends on organisational conditions. A recent BIS working paper examines factors associated with adoption of supervisory technology [18]. It supports considering staff capability, support and maintenance. Its observational results do not establish the benefits of the system proposed here.

2.4 Evidence provenance and access control

Earlier research already links information access to a person's task [24], the permitted purpose [7] and conditions that continue to apply during use [21]. W3C PROV provides a way to record where information came from, how it was derived and when its invalidation was recorded [27]. Doyle's work on "truth maintenance" records reasons for beliefs and revises those beliefs when their supporting assumptions change [14]. The present model uses a simpler check of evidence versions. It does not introduce a new theory of reasoning or automate supervisory judgement.

Related work also offers ways to show or hide parts of data according to the permitted purpose [25], combine access rules with agreed meanings [17]. Other studies address purpose-limited messaging [26], links between access rules and data history [28], and reusable rules and software for sharing data [10, 11, 22]. Research on regulatory technology also highlights the difficulty of turning a software design into demonstrated compliance [23].

Governance Inside the Agent Loop distinguishes what an actor can observe, what that actor is authorised to do and what actions the deployed system can carry out [12]. It also separates a previously valid approval from its applicability when

an action is taken. These distinctions inform the implementation requirements below. Its results concern a finite model of agent execution; they do not prove the usefulness of shared evidence in human regulatory supervision.

These sources establish the principal concepts underlying the proposed record. This paper applies them to a specific supervisory problem: identifying when earlier decisions need review and evaluating whether explicit evidence links improve reassessment. Its contribution is a defined model and a reproducible study design. It does not claim to be the first system to connect evidence, permissions and decisions.

3 A shared evidence record for supervision

The proposed record brings together a supervisory question, the evidence relevant to it, agreed definitions, access permissions, unresolved questions and the supervisor's decision. It also records which versions of the evidence and definitions were used. Each participant sees only the material they are authorised to see. "Shared" means that the organisations can refer to consistent evidence and meanings, not that everyone receives the same information.

For each item, the record identifies its source, when it applied, when it was received, how it was processed and any qualifications. When an item is corrected, the earlier version is retained and affected assessments are flagged for review. Changes to definitions or access permissions are also recorded. The system must distinguish an absence of evidence of a problem from evidence that the problem is absent.

For example, a firm may initially report that a service failure has been contained. Later information shows that other important services rely on the affected service. The supervisor must decide whether to ask further questions, monitor more closely or take stronger action. The shared record should connect the new information to the original report, earlier decision and outstanding corrective work. This is a constructed example, not an account of a real institution.

The supervisor remains responsible for the decision. The firm can provide evidence and challenge factual errors without gaining access to the regulator's internal discussions. An independent reviewer can contribute an assessment without receiving unrelated records. These arrangements preserve different responsibilities while allowing relevant evidence to be connected.

3.1 Evidential sufficiency

If situations cannot be distinguished using the evidence available to an authorised supervisor, yet have no response that is justified in all of them, that evidence cannot support a uniformly justified choice. Different sets of acceptable actions do not imply insufficiency when those sets have a common justified response. Possible remedies include an authorised summary, further enquiry or an explicit record that the assessment remains unresolved. Wider access to raw data is not always necessary or appropriate.

Many supervisory cases allow more than one reasonable response. Before a study begins, independent experts should agree which responses are acceptable and what reasoning each requires. Where the evidence is incomplete, a focused enquiry or a recorded decision to wait can be justified. Treating rapid closure as sufficient evidence of success would reward unsupported certainty.

3.2 Access, accountability and the historical record

Access must reflect the applicable authority and purpose. A regulator's right to obtain information does not necessarily depend on a firm's consent. Conversely, a technical connection does not itself create a right of access. Rules on retention, confidentiality, onward sharing and challenges to decisions must be checked in the jurisdiction where the system would operate. This paper does not establish legal powers.

Revoking access can prevent subsequent delivery but does not, by itself, remove information already received. Recording the history of evidence does not establish its veracity. New information must also leave intact the record of what was known when an earlier decision was made.

3.3 Linking decisions to their evidential basis

The model gives each evidence record four identifiers: its case, source version, interpretation version and permission version. The interpretation version identifies the approved rules for translating or comparing the information. The permission version changes when relevant access arrangements change. In the formal description these four identifiers are written as (c, v, m, p). Case identifiers are stable and unique. Each relevant version strictly increases and is never reused, including after a permission is renewed; otherwise an old decision could incorrectly become current again.

A full decision record would contain the decision time, action, authorised decision-maker, reasoning and these four identifiers. The executable model stores only time, action and evidence. It checks identity before accepting a record; a separate scoring function receives an externally supplied judgement about whether the reasoning is acceptable. Recording the full identity and reasoning history remains a deployment requirement. The study's event schedule records when information applies and when it arrives; the model itself handles receipt events in order.

A decision is marked as "current" only while its recorded identifiers match the latest evidence record and the relevant permission remains active. This establishes consistency with the latest recorded evidence and permissions, rather than substantive correctness. Earlier records are retained unchanged. Historical retrieval requires its own access policy; the

model's current-release check is not a policy for reading archived records.

A relevant correction to the source or its interpretation changes the corresponding version and flags the decision for review. A permission change prevents further release under the old permission version. It does not erase history or determine whether the original decision was substantively right. An unrelated event leaves the assessment unchanged. The model receives relevance classifications as inputs; it does not infer all dependencies among records in operational systems.

Provided each update completes before the next begins, a change to one identifier prevents an older decision from matching the latest record. This follows directly from the design, rather than being a new mathematical finding. Its effect on the timely identification and revision of outdated assessments is a separate empirical question.

Consider a correction that both systems receive immediately. One system links it to an earlier decision; another displays it without showing that connection. The information is equally fresh, but the need to reconsider the decision may be less apparent in the second system. However, evidence links cannot compensate for the absence of information necessary for a decision. These examples explain possible differences; they do not establish how often they occur or whether a particular existing system suffers from them.

3.4 A worked example

In one constructed case, an incident initially appears contained, so monitoring is reasonable. At the next assessment stage, a corrected report shows that an important service is seriously affected, making stronger supervisory action appropriate. Each stage has an illustrative 20-minute response window. Taking the required action five minutes after the correction counts as timely. Continuing to rely on the earlier monitoring decision does not meet the second stage's requirements.

If the participant later closes the case without adequate grounds, the decision at the deadline fails the quality check despite the earlier timely response. If the new evidence instead leaves the impact uncertain, a reasoned enquiry may be appropriate. These examples illustrate the scoring rules. Independent practitioners must validate them before participant testing; they are not instructions on how a regulator must act.

3.5 Information, authority and available action

Effective evidence sharing requires alignment between information availability, decision authority and the capacity to act. The person responsible for a decision must receive enough authorised evidence to justify it, have authority for the proposed response, and have a working means to carry out or refer that response. A firm's technical specialist may understand a service failure without having supervisory authority. A supervisor may have authority to act while lacking the specialist's evidence. Neither informational access nor decision authority alone is sufficient to establish a justified response.

An appropriate remedy may be an authenticated explanation or authorised summary sent to the responsible supervisor. It may instead be referral to a suitably informed and authorised person, or correction of the process through which the decision is carried out. Remedies should address the specific deficiency without presuming a general expansion of access or authority. This distinction [12] informs a design requirement; the associated formal guarantees do not extend to the present model.

3.6 Revalidation at the point of action

The evidence and authority supporting a proposed decision may change before that decision is implemented. For example, a supervisor may prepare to close an incident after receiving recovery evidence. Before closure is recorded or communicated, a correction arrives or the person's delegated authority changes. The system should check the relevant evidence, permissions and decision authority again at that point. If any premise has changed, it should require reconsideration rather than silently reuse the earlier approval.

The current model checks access when a decision is recorded and processes events one at a time. It does not implement a separate drafting-and-action sequence, delegated decision powers or concurrent updates. A deployed system must either complete its final check and action together or detect intervening changes and repeat the check. The checks proposed in Appendix B distinguish these future requirements from the tests already run.

Preventing reliance on an outdated decision should preserve an authorised next step, such as further enquiry, a reasoned deferral or referral to another reviewer. The study measures non-completion and unnecessary escalation, preventing universal refusal from satisfying its success criteria. These criteria do not guarantee eventual resolution in operational practice. Retrieval of similar cases may support assessment, but similarity alone does not justify adopting an earlier decision. Reuse requires verification of current evidence, definitions, permissions and authority.

4 Hypotheses and proposed mechanisms

H1: When teams receive the same authorised evidence at the same time, connecting that evidence to supervisory decisions reduces the average time to well-supported initial and revised decisions, compared with competent existing systems that organise the evidence separately. This reduction must satisfy predefined conditions for decision quality, combined workload and disclosure control.

The study compares independent teams across the same fixed set of assessment stages. Time is capped at each stage's deadline, as explained in section 5.3. The proposed benefit comes from reducing work spent finding evidence, resolving

differences in meaning or version, identifying missing information and reconstructing why an earlier decision was made.

The proposed mechanisms require empirical examination. Before registering the study plan, researchers must specify how logs and observations will measure retrieval and reconciliation work. Results for initial and revised assessments must also be reported separately: a pooled improvement driven entirely by initial decisions would not demonstrate better reassessment. To identify the value of one feature, such as linking a changed record to an earlier decision, a further experiment would need to vary that feature while holding the others constant. Implementation demands, excessive alerts or unwarranted confidence may offset the expected benefit.

H2: When evidence is organised in the same way, providing updates in response to relevant changes reduces the time between a change and a well-supported supervisory response, compared with the existing update process.

H2 concerns the timing of evidence delivery; H1 concerns evidence organisation. More frequent updates could create work without improving decisions. Equally, connected evidence could be useful where periodic reporting remains appropriate. A positive H1 result would support the capabilities tested, but would not establish the necessity of a particular platform or the value of every feature it contains.

5 Prospective evaluation design

5.1 Comparison conditions

The study proposes three arrangements, where the participating regulator's existing systems allow a meaningful comparison. C is current practice, including its actual reporting schedule, notifications, links between records and analytical tools. U delivers updates in response to relevant events but keeps the existing way of organising evidence. G uses the same update schedule and access rights as U, while connecting evidence and decisions in the shared record described above.

The main comparison, G against U, tests H1. Comparing U with C tests H2 only if the intended difference is the delivery schedule. Comparing G with C tests the combined change in delivery and organisation.

The comparator must represent existing capabilities accurately. No group should repeat approvals or perform manual tasks that its normal tools already automate. G and U must have equivalent ways to transform and analyse the information, as well as equivalent generic arrival notifications, training and support. Dependency-specific review flags are part of the G intervention and must be documented as such. Differences in the effort needed to find and reconcile evidence are part of what the study aims to measure. Differences in information content or analytical capability would confound the intended comparison.

The intended differences between conditions must be verified before participant allocation. If current practice already connects evidence and decisions in the proposed way, there may be no integration difference to test. If changing delivery times also changes the analysis available, the U–C comparison tests a package of changes, not timing alone. Artificial restrictions on the comparator would invalidate conclusions about improvement over existing practice.

5.2 Cases and participants

The supplied examples cover six situations: corrected information, recovery without further problems, missing verification, conflicting independent assessments, a recurring incident and a changed interpretation. Each case has two assessment stages, initially allowing 20 minutes per stage. Timing may be revised after the pilot but before the main study is registered. Participants cannot act before a stage opens. The first stage closes before the next evidence release, and every response must identify its stage. Both stages run to their deadlines even when participants respond early.

The first study should focus on operational incidents and the work undertaken to resolve them. This provides changing evidence and interaction between firms and supervisors without attempting to reproduce every regulatory activity. Other areas, such as financial soundness or business conduct, require later studies.

Independent supervisory and industry practitioners must develop and validate the cases before testing. They should include genuine deterioration, harmless changes, unresolved questions and claims of recovery that lack supporting evidence. The same case developments must be presented to each group. A panel independent of the system's developers should agree the response windows, serious-error definitions, acceptable actions and required reasoning. More than one judgement may qualify. Disagreements must be recorded and resolved through a process agreed in advance.

An initial pilot would involve 24 independent teams, 12 using G and 12 using U. The pilot is intended to assess procedural and measurement feasibility rather than effectiveness. Training uses separate cases, and pilot participants do not take part in the main study. A C group would be planned separately and treated as exploratory until the timing comparison is verified.

Teams would be assigned randomly to G or U within groups of similar experience, and stay with that arrangement to reduce learning across conditions. The team remains the independent unit of analysis across repeated cases. If shared staff, firms or workflows allow one team's participation to affect another's, the assignment and analysis must be redesigned around a larger independent unit. Study participation must not influence real supervisory decisions.

5.3 Outcome measures

For H1, elapsed time is measured from the scheduled beginning of each assessment stage, identically for G and U. This applies even when missing evidence means that the justified response is to ask a question. Timing must not depend on when

a participant finds an item or when all uncertainty disappears.

The response-time measure is the interval to the first justified response, capped at the stage deadline. Where no justified response occurs, the full interval is recorded alongside non-completion. A response exactly at the deadline and no response therefore receive the same time value but different completion results. In unresolved cases, an appropriate enquiry or recorded decision to wait can qualify; an unsupported closure cannot.

For H2, also measure the full time from the scripted change to the justified response. Separate the delay in delivering information from the time taken to assess it. Decisions are recorded throughout both stages so that an acceptable early response does not obscure a subsequent serious error.

Quality measures cover missed deterioration, unjustified closure, unnecessary stronger action and appropriate handling of uncertainty. For the main statistical analysis, completion means that a team gives a justified response by every stage's deadline. Final quality means that its decision at the end of every stage remains justified. Separately record whether the team makes any serious error, takes any unnecessary stronger action, or releases any information without permission anywhere in the six cases. These five measures are binary outcomes assessed for each team across the fixed set of cases. Also report results by stage, but do not treat stages as independent teams.

Both groups use the same response form. Reviewers compare responses with an independently held schedule of the case events and evidence. They judge whether the response addresses the relevant facts and uncertainty, not whether it uses G's version numbers or record format. A participant may reasonably reaffirm an earlier decision after considering a change. The model's "current" status is distinct from the independent assessment of human decision quality. Cases should include changes that do not warrant escalation, allowing the evaluation to identify indiscriminate escalation. Outcome reviewers should be blinded to team allocation where feasible; participants cannot be blinded to the interface they use.

Combined workload is the sum of active work measured separately for supervisors, firm staff, independent reviewers and support staff. Count preparation, clarification, assessment, follow-up and time spent dealing with failures within each observation window. Use person-minutes: four people working for 20 minutes contribute 80 person-minutes. Average this across the same 12 stages used for the timing measure. The illustration assumes no more than four people working concurrently; the actual staffing limit must be agreed before the study.

Preparation outside those windows, setup, training, maintenance of agreed definitions and ongoing operating costs belong in a separate cost assessment. Acceptable workload during the exercise does not establish lower total regulatory burden. A reduction in supervisory workload that increases firms' workload may not produce a net benefit.

Distinguish unauthorised disclosure from collecting more information than necessary despite having permission. Separately test expired permissions, access to another case's records, corrections and failures in the audit record. A small study cannot show that rare confidentiality failures are impossible. If none is observed, the analysis must still allow for uncertainty about the underlying risk.

5.4 Statistical analysis and decision criteria

The statistical plan requires G to meet all seven conditions: a practically meaningful reduction in average capped response time; no unacceptable increase in combined work during the observation windows; no unacceptable decline in completion or final quality; and no unacceptable increase in the risks of serious error, unnecessary stronger action or unauthorised disclosure. Practitioners must agree the minimum practically meaningful benefit and maximum acceptable losses before the main study. Improvement over a comparator is insufficient for adoption unless the proposed system also meets independent minimum standards and implementation requirements.

The analysis compares team averages, allowing for experience groups using weights fixed in advance. It reports uncertainty around the differences rather than relying on averages alone. For time and workload, the method uses the fact that observations have agreed upper and lower limits. For binary outcomes, it uses binomial confidence limits that retain uncertainty even when no adverse events are seen.

For reproducibility, the technical methods are one-sided Hoeffding bounds for bounded team observations [20] and exact Clopper–Pearson limits within each study group and experience group [8]. The latter are combined using a union bound to obtain limits for the difference in event probabilities. Appendix A gives the formulae and decision directions; the supplementary statistical analysis plan specifies implementation and constructed simulation inputs. These methods require independent teams; the binomial calculation also assumes the same event probability for teams within each study and experience group. The target is the weighted mean or event-risk difference in the specified team populations, conditional on the fixed case set and stratum sample sizes. It is not a distribution-free randomisation interval for a fixed finite roster. Responses within a team can be related. Influence between teams would violate the independence assumption.

Each one-sided bound has at least 97.5% coverage under these assumptions: over repeated comparable studies, the procedure would place the bound on the appropriate side of the true difference at least that often. This is not a 97.5% probability that a particular reported bound is correct. Requiring all seven conditions to pass is an intersection–union test at level 0.025 for the single joint claim: if any required condition is false, a full pass entails rejection of that true component null. This does not require independence across outcomes or a sevenfold adjustment, and does not provide simultaneous 97.5% coverage for all seven bounds. A reduction in response time cannot compensate for failure to meet a quality or workload condition. Pairing the lower and upper one-sided bounds gives at least 95%, not 97.5%, two-sided coverage. Separate claims about H2 or G against C remain exploratory unless registered with an analysis that accounts for testing additional claims.

Missing observations are handled conservatively. All assigned teams remain in the analysis. For a missing binary outcome, the calculation allows either possible value; for missing time or workload, it allows the full range consistent with what is known. The reported bounds cover those possibilities. Work exceeding the agreed staffing limit must be reported and invalidates the unamended workload bound; such observations must not be truncated to preserve the assumed range. Teams that abandon the exercise and outages caused by the tested system remain part of the results.

The supplied simulations examine how often all seven conditions would pass together under stated assumptions. They do not select the number of teams for a real study. The pilot must establish whether recruitment, realistic cases and reliable observation are feasible. It cannot reliably estimate very rare confidentiality failures. If a study capable of testing the full claim is impractical, report feasibility findings or estimates with their uncertainty, or validate a more efficient analysis before proceeding. Success criteria must not be relaxed after inspecting results, and repeated requests within a team must not be counted as independent observations.

6 Implementation requirements and model evaluation

The proposed approach requires agreed meanings for the information being compared; access matched to roles and purposes; records of evidence versions and dates; visible gaps and uncertainty; links from relevant updates to earlier assessments; and a record of the reasons for each decision. These capabilities can be provided by different technical designs. No particular vendor, central database, distributed ledger or language model is required.

Each capability requires explicit verification criteria. If a report changes which companies it covers, affected comparisons should be identified. A corrected submission should leave the earlier version and decision history accessible to authorised reviewers. A person without permission should not receive protected evidence. An unresolved discrepancy should remain visible when a decision is recorded. Passing these checks shows that specified functions work, not that supervision has improved.

Initial implementation should address a defined recurring assessment and a limited set of agreed terms. Firms should be able to provide evidence through existing processes where feasible. Evaluation should include setup and maintenance effort for both small and large firms. Persistent dependence on specialist support may undermine cost-effectiveness despite improvements in supervisory usability.

6.1 Executable model and verification results

The executable model processes events one at a time. It keeps earlier evidence unchanged, records an audit entry before releasing information and links decisions to the versions they used. Changes to sources, interpretation rules or permissions trigger a review requirement.

The model tests all 64 combinations of six request conditions, of which one permits access, and checks the audit outcome for each combination. It also tests second-stage scoring examples for the six two-stage cases; it does not run a complete participant workflow through both stages. It passes 184 checks. To examine whether the checks detect faults, eight protections were removed in turn: identity, case identity, source version, interpretation version, permission version, audit availability, use of current evidence and prevention of a backwards-moving clock. Each deliberately introduced fault was detected. This is selected test coverage, not a complete security assessment.

The model trusts the supplied identity, clock, relevance judgements and scoring inputs. It does not establish legal authority, verify source truth, discover every affected decision, prevent disclosure through inference, implement timed permission expiry, or handle simultaneous operations across multiple systems. Its invariants assume use of the controlled update interface; Python object access is not a security boundary. Its audit record is held in computer memory rather than a durable storage system. Information already released remains readable after access is withdrawn. These limitations preclude interpreting the results as evidence of improved human decision-making or production readiness.

The scoring examples distinguish an early justified response from the decision retained at the deadline. They reject reasoning marked as using outdated evidence and distinguish a response at the deadline from non-completion. There is no participant interface. Before recruitment, a separate check must establish that G and U provide equivalent information, permissions, analysis and release times, without artificial delays.

6.2 Statistical verification and planning simulations

The statistical implementation passes 32 checks. These include enumerating all possible event counts in 18 binomial settings and checking how often a component confidence bound would fail. The largest calculated failure probability is 0.0112532, below its allowed 0.0125. These examples check the implementation. The general mathematical justification comes from the published methods and the assumptions stated above.

The planning simulation repeats each setting 400 times, using five sample sizes per group and five scenarios. Under its illustrative benefit assumptions, all seven conditions pass in 0% of simulations with 48, 96, 192 or 384 teams per group, and 82.75% with 768. Scenarios with no practically meaningful time benefit, worse decision quality or extra combined workload produce no full passes at these sample sizes. A scenario with independently missing observations at a probability of 2% per team, treated conservatively, also produces no full passes. These results apply to the accompanying reference implementation and its stated synthetic inputs.

Sampling variation from running a finite number of simulations gives a standard error of at most 2.5 percentage points. An observed 0% or 100% is not proof that the exact probability is zero or one. These results use synthetic effect sizes and thresholds; they are neither estimates of regulatory benefits nor recommendations for study sample size. Separate standards for actual deployment are not included. The simulations demonstrate the limitations of planning solely for a response-time improvement and the importance of recruitment feasibility and complete observation when assessing the joint hypothesis.

6.3 Failure diagnosis and corrective measures

Failure analysis must distinguish the cause of an unsuccessful response. Missing evidence calls for enquiry or an authorised summary; missing decision authority calls for a valid referral; an outdated approval calls for a fresh check; and an unavailable action requires a process or system repair. These are different interventions. Appendix B separates the corresponding checks so that a failure in one layer is not misreported as proof that all evidence sharing is ineffective.

7 Further research and limitations

Following practitioner validation and the initial comparison, the approach could be evaluated alongside operational workflows without influencing live supervisory decisions. Comparable evidence should be provided to the study teams, and any information they exchange must be recorded. Duplicated activity required by the study should be recorded as a research cost rather than interpreted as an operational saving. Actual adoption requires a separate assessment of authority, security, accountability, organisational incentives and workload.

This paper combines a critical literature review, a bounded executable model and a prospective evaluation design. It does not establish improved human performance or readiness for live use. Institutional reports help explain the problem but cannot replace comparative observations. Constructed cases may simplify genuine disagreements. Participant experience, access to firms and organisational incentives may change the results. Findings about operational incidents would not automatically apply to market surveillance, financial crime or all forms of financial supervision.

The research framing originates in work connected to the development of Pathiqa, a commercial decision-support platform whose Views capability is a feature rather than a separate product. This commercial connection is a potential competing interest. No product implementation has been evaluated in this paper. Any platform used in a later study must be identified, with independent control over outcome assessment and publication of unfavourable results. No regulator's involvement or endorsement is implied.

8 Conclusion

Linking changes in evidence to the supervisory decisions that depend on it provides a coherent basis for continuous supervision. The central proposition is that improvements in supervisory responsiveness depend not only on the timeliness of information, but also on the ability to identify which assessments require reconsideration and why. A shared evidence record provides an explicit mechanism for maintaining these relationships while preserving access restrictions, uncertainty and the historical basis of decisions.

The expected benefits arise through reduced evidence retrieval and reconciliation, earlier identification of assessments requiring review, and a clearer evidential basis for subsequent action. These mechanisms may improve supervisory performance even when information content and delivery times remain unchanged. Their net value depends on whether these gains outweigh implementation costs, additional alerts and any increase in firms' workload.

The paper's contribution is a specified mechanism, verification of selected model properties and a reproducible design for comparative evaluation. It establishes a testable basis for the proposed approach; it does not establish improvements in human decision-making, a reduction in total regulatory burden or effectiveness in operational supervision. These outcomes require participant and field evidence.

Independent evaluation should therefore compare the proposed capabilities with existing practice under equivalent evidence availability and access rights. Support for the primary hypothesis requires a practically meaningful reduction in response time while satisfying predefined conditions for quality, completion, workload and disclosure. Such evidence, combined with a broader cost assessment, would support investment in the capabilities evaluated. An adverse or inconclusive result would delimit the conditions under which their value has been demonstrated. The argument concerns these capabilities rather than the necessity of any particular platform.

Supporting materials and evidence status

The supplement contains six constructed case outlines, a draft operational protocol, the statistical plan, a source register, a reference implementation, verification checks, targeted mutation tests and synthetic simulation outputs. The model uses Python's standard library; the statistical checks and simulations require NumPy and SciPy. The package includes execution instructions, dependency versions, the random seed and results for every simulation replicate. The outlines are not validated participant materials, and the protocol is not a registered study.

No human observations or confidential regulatory records are used. Timings, acceptable case responses and simulation thresholds are constructed inputs that require independent calibration.

A Model and assumptions

The shared record uses four identifiers: case c, source version v, interpretation version m and permission version p. Let $K = (c, v, m, p)$. A recorded decision D retains the identifiers K(D) that applied when it was made. Its model status is current only if K(D) equals the latest K and the relevant access permission remains active. This necessary check does not establish that D is correct or that its maker has legal authority to take every possible action.

Under sequential updates and non-reused versions, changing any component of K prevents an earlier record from passing the equality check. This is a property of the definition, not a new theorem. Historical evidence and decisions remain unchanged. The implementation compares the stored evidence snapshot; its controlled update functions advance the relevant version when they change the snapshot.

Table A1. Assumptions and what they leave unresolved.

Assumption	Why it is needed	What remains outside the guarantee
Identity, time and incoming evidence are trusted	Access and ordering depend on these inputs	Compromised sources and false statements
Relevant changes are identified externally	Relevance inputs determine when review is required	Discovering all affected decisions
Updates occur one at a time	No change can interleave with the model's operation	Concurrent systems and action-time races
Case reasoning is judged externally	A version match cannot assess a human argument	Expert agreement and legal correctness
Study teams are independent	Confidence calculations use independent team observations	Shared-staff and organisational interference
Work and time remain in declared ranges	Bounded statistical methods require fixed limits	Unrecorded work or excess staffing

For the proposed study, define each team's time as its average across the same 12 assessment stages, with each stage capped at 20 minutes. Work is the corresponding average of active person-minutes, initially bounded by 80 per stage. Both limits require confirmation before registration. Within each predeclared experience group s, compare the G and U means. Use positive weights w_s summing to one, fixed before examining the results.

The estimated difference is

$$\widehat{D} = \sum_s w_s(\overline{X}_{Gs} - \overline{X}_{Us}).$$

With positive arm–stratum sample sizes n_{Gs}, n_{Us} and independent team observations in $[0, R]$, the one-sided Hoeffding radius is

$$r = R\sqrt{\frac{\log(1/\alpha)}{2}\sum_s w_s^2\left(\frac{1}{n_{Gs}} + \frac{1}{n_{Us}}\right)}, \qquad \alpha = 0.025.$$

The bounds $\widehat{D}-r$ and $\widehat{D}+r$ are each one-sided limits; together they have at least 95% coverage. For binary outcomes, the plan uses one-sided Clopper-Pearson limits at $\alpha/(2S)$ within each group, where S is the number of experience groups. Combining the relevant $2S$ component bounds gives each direction of the weighted risk-difference bound. The supplementary statistical analysis plan specifies missing observations, decision thresholds and assumptions; these are established statistical methods, not new results [8, 20].

For k events among n teams and tail level $\beta = \alpha/(2S)$, write $Q_\gamma(a, b)$ for the γ quantile of a Beta(a, b) distribution. The component limits are

$$L(k,n) = \begin{cases} 0, & k = 0, \\ Q_\beta(k, n-k+1), & k > 0, \end{cases} \quad U(k,n) = \begin{cases} 1, & k = n, \\ Q_{1-\beta}(k+1, n-k), & k < n. \end{cases} \tag{A.1}$$

Consequently, one-sided limits for the weighted risk difference $D = \sum_s w_s(p_{Gs} - p_{Us})$ are

$$L_D = \sum_s w_s(L_{Gs} - U_{Us}), \qquad U_D = \sum_s w_s(U_{Gs} - L_{Us}). \tag{A.2}$$

Each direction fails only if at least one of its $2S$ component limits fails, so its failure probability is at most $2S\beta = \alpha$. No independence between these component limits is needed for this union bound.

Let $\delta_T > 0$ be the required time benefit and let $m_W, m_C, m_Q, m_E, m_N, m_D > 0$ be acceptable losses. All differences are G minus U. The joint claim requires

$$U_T < -\delta_T, \qquad U_W < m_W, \qquad L_C > -m_C, \qquad L_Q > -m_Q, \tag{A.3}$$

$$U_E < m_E, \qquad U_N < m_N, \qquad U_D < m_D. \tag{A.4}$$

Here C, Q denote completion and final quality (higher is better); E, N, D denote serious error, unnecessary escalation and disclosure (lower is better). Equality at a threshold does not pass. Non-significance is not evidence of non-inferiority.

For missing observations, retain the assigned denominators. Maximise the estimated bounded difference by assigning missing G values R and missing U values zero; reverse these assignments for its minimum, then apply the same radius. For a binary cell with k observed events and h missing outcomes among n assigned teams, use $L(k, n)$ and $U(k + h, n)$. These envelopes include the complete-data bounds regardless of the missingness mechanism, provided the outcome ranges and assigned denominators are known.

B Failure checks and evidence status

Table B1. Completed checks and additional checks required before field use.

Failure case	Required response	Evidence status
Wrong identity, case or evidence version	Refuse the request	Tested in the sequential model
Source, interpretation or permission changes	Retain history; flag the previous decision for review	Tested in the sequential model
Audit unavailable or clock moves backwards	Refuse the affected release or decision record	Tested in the sequential model
Initial justified response followed by unsupported closure	Retain early completion but fail final quality	Tested with scoring examples
Evidence changes after drafting but before action	Repeat the final check or require reconsideration	Required for deployment; not implemented
Informed person lacks decision authority	Refer evidence to an informed, authorised decision-maker	Requires practitioner validation and workflow testing
Prior decision retrieved from a similar case	Check present evidence, meaning, permissions and authority	Required for deployment; not a tested reuse engine
Checks prevent every response	Record non-completion; identify a lawful next step	Outcome defined; human workflow not tested

The first four rows summarise existing checks; they do not cover every failure of their type. The later rows are proposed requirements and must not be counted as completed experiments. The eight deliberately introduced code faults and exact results are recorded in the reproducibility supplement. Neither these checks nor the formal results in [12] establish the suitability of a production implementation.

C Terminology

Table C1. Terminology used in the manuscript and technical supplement.

Term in this paper	Technical term in the supplement	Meaning
Shared evidence record	Governed supervisory evidence context	Linked evidence and decisions with controlled access
Assessment stage	Checkpoint	A fixed period in which a response is assessed
Well-supported response	Justified disposition	An acceptable action with the required reasoning and authority
Final quality	Terminal quality	Whether the decision retained at each deadline is justified
Interpretation version	Mapping version	Which approved definitions or translation rules were used
Permission version	Authorisation epoch	Which access arrangements applied; not proof of decision authority
Stronger supervisory action	Escalation	A more intensive response, as defined in the validated case
Joint satisfaction of all decision criteria	Intersection-union test	A gain on one required outcome cannot offset failure on another

References

[1] Raphael Auer. *Embedded supervision: how to build regulation into decentralised finance*. BIS Working Papers 811, Bank for International Settlements, September 2019. URL https://www.bis.org/publ/work811.htm.

[2] Henrik Axelsen, Johannes Rude Jensen, and Omri Ross. DLT Compliance Reporting. arXiv:2206.03270v1, May 2022. URL https://arxiv.org/abs/2206.03270v1. Preprint; submitted 31 May 2022.

[3] Kenton Beerman, Jermy Prenio, and Raihan Zamil. *Suptech tools for prudential supervision and their use during the pandemic*. FSI Insights 37, Bank for International Settlements, December 2021. URL https://www.bis.org/fsi/publ/insights37.pdf.

[4] BIS Innovation Hub. *Project Ellipse: An integrated regulatory data and analytics platform*. Project report, Bank for International Settlements, 31 March 2022. URL https://www.bis.org/publ/othp48.pdf.

[5] Julia Black and Robert Baldwin. Really Responsive Risk-Based Regulation. *Law & Policy*, 32(2):181–213, 2010. doi:10.1111/j.1467-9930.2010.00318.x.

[6] Dirk Broeders and Jermy Prenio. *Innovative technology in financial supervision (suptech)—the experience of early users*. FSI Insights 9, Bank for International Settlements, July 2018. URL https://www.bis.org/fsi/publ/insights9.pdf.

[7] Ji-Won Byun and Ninghui Li. Purpose based access control for privacy protection in relational database systems. *The VLDB Journal*, 17(4):603–619, 2008. doi:10.1007/s00778-006-0023-0.

[8] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi:10.1093/biomet/26.4.404.

[9] Juan Carlos Crisanto, Katharina Kienecker, Jermy Prenio, and Eileen Tan. *From data reporting to data-sharing: how far can suptech and other innovations challenge the status quo of regulatory reporting?*. FSI Insights 29, Bank for International Settlements, December 2020. URL https://www.bis.org/fsi/publ/insights29.pdf.

[10] Tobias Dam, Lukas Daniel Klausner, Sebastian Neumaier, and Torsten Priebe. A Survey of Dataspace Connector Implementations. arXiv:2309.11282v2, January 2024. URL https://arxiv.org/abs/2309.11282v2. Version consulted revised 9 January 2024; associated with ITADATA 2023 proceedings.

[11] Tobias Dam, Andreas Krimbacher, and Sebastian Neumaier. Policy Patterns for Usage Control in Data Spaces. arXiv:2309.11289v1, September 2023. URL https://arxiv.org/abs/2309.11289v1. Preprint; submitted 20 September 2023.

[12] Decision Superintelligence Labs. *Governance Inside the Agent Loop: Exact Nonblocking Control under Partial Observation and Dynamic Authority*. Research paper, 1 August 2026. URL https://www.dsilabs.ai/research/governance-inside-the-agent-loop.

[13] Digital Regulatory Reporting team. *Digital Regulatory Reporting: Phase 2 Viability Assessment*. Report hosted by the Financial Conduct Authority, n.d.. URL https://www.fca.org.uk/publication/discussion/digital-regulatory-reporting-pilot-phase-2-viability-assessment.pdf.

[14] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. doi:10.1016/0004-3702(79)90008-0.

[15] European Central Bank. *ECB Annual Report on supervisory activities 2024*. March 2025. URL https://www.bankingsupervision.europa.eu/press/other-publications/annual-report/html/ssm.ar2024~700cba1314.en.html. Section 5.2.

[16] European Central Bank. *ECB Annual Report on supervisory activities 2025*. 2026. URL https://www.bankingsupervision.europa.eu/press/other-publications/annual-report/pdf/ssm.ar2025~6ee989dc7e.en.pdf. Section 7.1.

[17] Nikos Fotiou, Vasilios A. Siris, and George C. Polyzos. Access control for Data Spaces. *Proceedings of ICIN 2025*, pages 54–58, 2025. doi:10.1109/ICIN64016.2025.10943024. URL https://arxiv.org/abs/2504.13767v1. Author manuscript, arXiv:2504.13767v1, 18 April 2025.

[18] Leonardo Gambacorta, Nico Lauridsen, Samir Kiuhan-Vásquez, and Jermy Prenio. *Making suptech work: evidence on the key drivers of adoption*. BIS Working Papers 1309, Bank for International Settlements, 25 November 2025. URL https://www.bis.org/publ/work1309.htm.

[19] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004. URL https://aisel.aisnet.org/misq/vol28/iss1/6/.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi:10.1080/01621459.1963.10500830.

[21] Jaehong Park and Ravi Sandhu. The UCON$_{ABC}$ usage control model. *ACM Transactions on Information and System Security*, 7(1):128–174, 2004. doi:10.1145/984334.984339.

[22] Jérôme Pfeiffer, Nicolai Maisch, Sebastian Friedl, Matthias Milan Strljic, Armin Lechler, Oliver Riedel, and Andreas Wortmann. Declarative Policy Control for Data Spaces: A DSL-Based Approach for Manufacturing-X. arXiv:2511.22513v1, November 2025. URL https://arxiv.org/abs/2511.22513v1. Preprint; submitted 27 November 2025.

[23] Paul Ryan, Martin Crane, and Rob Brennan. Design Challenges for GDPR RegTech. *Proceedings of ICEIS 2020*, 2020. URL https://arxiv.org/abs/2005.12138v1. Author manuscript, arXiv:2005.12138v1, 21 May 2020.

[24] R. K. Thomas and R. S. Sandhu. Towards a task-based paradigm for flexible and adaptable access control in distributed applications. *Proceedings of the New Security Paradigms Workshop*, pages 138–142, 1993. URL https://www.nspw.org/papers/1993/nspw1993-thomas.pdf.

[25] Khai Tran, Sudarshan Vasudevan, Pratham Desai, Alex Gorelik, Mayank Ahuja, Athrey Yadatore Venkateshababu, Mohit Verma, Dichao Hu, Walaa Eldin Moustafa, Vasanth Rajamani, Ankit Gupta, Issac Buenrostro, and Kalinda Raina. Data Guard: A Fine-grained Purpose-based Access Control System for Large Data Warehouses. arXiv:2502.01998v2, October 2025. URL https://arxiv.org/abs/2502.01998v2. Preprint; version consulted revised 21 October 2025.

[26] Karl Wolf, Frank Pallas, and Stefan Tai. Messaging with Purpose Limitation—Privacy-Compliant Publish-Subscribe Systems. arXiv:2110.15150v1, October 2021. URL https://arxiv.org/abs/2110.15150v1. Preprint; submitted 28 October 2021.

[27] World Wide Web Consortium. *PROV-DM: The PROV Data Model*. W3C Recommendation, 30 April 2013. URL https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[28] Faen Zhang, Xinyu Fan, Wenfeng Zhou, and Pengcheng Zhou. Purpose-based access policy on provenance and data algebra. arXiv:1912.00445v1, December 2019. URL https://arxiv.org/abs/1912.00445v1. Preprint; submitted 1 December 2019.