

Metric Baselines & Signal Validation

A reproducible monthly number can still be too immature to support its conclusion

A recurring performance-reporting case that separates source version, event cohort, maturation, goal basis and action evidence. It tests the value of resolving uncertainty before publishing a direction.

MONTHLY PERFORMANCE & EVIDENCE GOVERNANCE

BLUF / BOTTOM LINE UP FRONT

Lock what was measured. Qualify what the monthly change can establish.

ANALYTICAL FINDING

A version lock prevents silent substitution; it does not establish truth or equal capture. Incomplete monthly cohorts can support improvement, no change or deterioration under the same observed comparison.

CONTRIBUTION IN THE SOURCE RECORD

Provisional registry-level hoop for the reporting artifact, its specification and number-basis decisions. Operating results remain separate.

EVIDENCE STATUS

Exact monthly reporting and decision originals remain unrecovered. A related planning artifact cannot authenticate release, execution or impact.

BLUF: analytical inference. Source contribution: historical record. Client and internal identities anonymized.

STRUCTURED ABSTRACT

Release the monthly decision with its evidence limits intact

[AI] A reproducible report can preserve the wrong basis perfectly. The operating decision needs a stable record and a defensible interpretation of what that record measures.

Element	Decision-relevant conclusion
Context [HR]	A planning lead reportedly developed a recurring performance report, a root-cause/action matrix, a locked-number spine and executive discussion materials. The record supports a reporting-design case; exact originals, release and adoption remain unauthenticated.
Objective [AI]	Separate source version, original cohort, event maturity, target authority and action evidence before using a monthly movement to change priorities or credit programmes.
Approach [RC/AI]	Reconcile rival versions and scenario arithmetic privately. Test original-cohort uncertainty and distinguish descriptive labels from causal mechanisms. Evaluate a finite, explicitly invented shadow reporting workflow with strong cheaper alternatives.
Findings [AI]	A coherent alternative baseline can explain several printed calculations without authenticating that baseline. A locked apparent decrease can permit either mature direction. A selected tail scenario does not bound every feasible tail intervention.
Decision value [PD]	The measurement owner signs the cohort and maturity basis; the goal owner signs the target version; the forum owner releases a qualified conclusion when admissible bounds cross the decision threshold. Operations separately approves and evaluates actions.
Implications [PD]	Apply the Vintage-to-Decision Evidence Contract (VDEC). The tested projection improves on an unqualified headline but loses to a free qualified report. More complete tracing cannot rescue a poorly chosen reporting rule.

[PD] Start with a version bridge, an original-ID resolution audit and an agreed error-cost policy. The proposed benchmark measures reporting decisions; it neither changes physical outcomes nor demonstrates historical savings.

[HR] historical record; [RC] recalculated; [AI] analytical inference; [PD] proposed design. Source quantities and identities remain private. Independent numerical constructions are explicitly hypothetical.

EXHIBIT 1 / THE BASIS CONTRACT

Freeze the source and the interpretation as separate objects

[AI] A number lock prevents silent substitution. Accuracy, comparability and decision authority require additional evidence.

Locked object	Question the lock does not settle
Source and pull version	Was the source complete, correctly linked and applicable to this cohort?
Event and observation clocks	Are prior and current periods equally mature under comparable capture?
Metric and population	Are numerator, denominator, exclusions and hierarchy definitions unchanged?
Goal and planning basis	Which authority approved this target for this population and period?
Report and action version	Was the report released, and was the recommended action accepted or executed?

[HR/AI] The reporting specification distinguishes source-of-truth figures, rival goal bases, missing inputs and prior-period material. That is useful governance. A designated source can still contain late revisions, conflicting populations or an internally inconsistent subtotal. An authoritative choice of goal does not make a competing actuals series comparable.

[PD] Preserve the initial release and append a restatement with its reason, affected cells and decision consequence. Never overwrite the prior comparison without a bridge. A reviewer should be able to reconstruct both the original decision information and the later corrected interpretation.

QUALIFY A CONCLUSION WITHOUT FREEZING THE OPERATION

[PD] A qualified report can publish verified facts, unresolved ranges, a responsible follow-up owner and a release deadline. It need not withhold existing customer remedies, fabricate a current value or wait indefinitely for certainty.

[HR/AI] A recurring reporting specification is evidence of design intent. Its existence does not prove that every source check ran, every exception was resolved or the forum adopted the output.

EXHIBIT 2 / ORIGINAL-COHORT UNCERTAINTY

A locked decrease can still conceal either mature direction

[PD] An independent construction shows why confirmed events and unresolved original IDs must travel together through the comparison.

Invented monthly frame	Prior month	Current month
Original eligible units	1,000	1,000
Confirmed eligible defects	80	60
Unresolved original statuses	20	80
Possible mature defect count	80 to 100	60 to 140
Observed comparison	Confirmed counts only	20 fewer
Mature current-minus-prior change	Joint uncertainty	-40 to +60

[PD] Confirmed labels are correct, each unresolved unit can contribute at most one eligible defect, and no eligible units lie outside the frame. Under these assumptions the bounds are sharp: every endpoint is attainable. They are logical identification bounds, not statistical confidence intervals. A different linkage, false-positive or event-multiplicity model requires different bounds.

[AI] Equal extraction age does not ensure equal resolution. A programme, contact channel, taxonomy or operational disruption can change who becomes visible. A faster current feed can look worse despite unchanged physical experience; a slower feed can look better. Freezing a file does not freeze that observation mechanism.

KEEP UNKNOWN STATUS IN THE ORIGINAL DENOMINATOR

[PD] Track unresolved IDs and resolution by event type, cohort and age. Use independent adjudication and show how remaining uncertainty changes the decision. Do not discard unobserved cases or turn a missing field into a verified negative.

[PD] Wholly invented two-cohort example, independently recalculated. It is separate from the later model, whose prior cohort is assumed fully verified and closed.

EXHIBIT 3 / RECONCILIATION BEFORE INTERPRETATION

A coherent version mismatch calls for a bridge, not selective correction

[RC/AI] Related arithmetic discrepancies can reveal a common hidden basis. Reconstructing that basis is a diagnostic result, not proof that it was the approved source.

Private-source finding	Public decision consequence
Several scenario calculations share an inferred alternative baseline	Seek the original version and preserve both calculations; do not repair cells independently without lineage.
Alternative prior values reverse the direction of a monthly comparison	Direction remains version-dependent until the source, cohort and maturity bridge is accepted.
A purported target beat remains above the stated target	Correct the direction and identify the applicable target; arithmetic cannot choose authority.
Reported component value exceeds the whole if their scopes match	Reconcile population, period, valuation and overlap before adding or attributing dollars.

[AI] A numerical identity can be internally consistent while relying on the wrong cohort or goal. Conversely, a difference need not be a calculation error if the records intentionally use different definitions. The audit first states the equality that would have to hold under matched scope, then checks whether its prerequisites are documented.

[PD] Store each source value with metric, population, event window, pull vintage, maturity and authority. A replacement value must name the old value and its reason. When the bridge cannot be recovered, publish the conditional interpretations instead of averaging incompatible versions.

DO NOT CONFUSE RECONSTRUCTION WITH AUTHENTICATION

[AI] The recovered arithmetic pattern supports a version-mismatch hypothesis. It does not establish which file was released, why the version changed or whether a downstream decision used it.

[RC/AI] All historical arithmetic and source coordinates are retained privately. The original monthly artifacts were not recovered; the public finding is deliberately qualitative.

EXHIBIT 4 / DIAGNOSIS AND ACTION

Use the action matrix to organize hypotheses, then test the mechanism

[AI] A defect label, a probable capture anomaly and a named playbook are different evidence states. None alone establishes the cause or effect of an intervention.

Reporting observation	Boundary on the inference
Abrupt change in a recorded defect	A data fault, real operating change and their combination remain possible without an independent reference.
Policy-adjusted metric	Mechanical eligibility or classification adjustment is not the no-policy causal counterfactual.
Low explicitly coded external-cause share	Uncoded exposure and indirect pathways can remain; absence of a label does not exclude the mechanism.
Fixed root-cause/action lookup	A candidate response still needs local applicability, accepted ownership, execution and mature outcomes.

[HR/AI] The mechanism used a stable narrative spine and an action bank to make recurring reports usable. That can reduce unstructured discussion. It can also repeatedly reproduce an untested causal label. A low-level taxonomy should not silently become a diagnosis, and a named programme should not acquire impact credit merely because it appears next to a declining metric.

[PD] Separate observed defect, proposed mechanism, discriminating evidence, authorized response and outcome evaluation. Quarantining a suspect series may be prudent; the quarantine itself does not prove the failure was exclusively in the reporting pipeline.

RETAIN BOTH PHYSICAL AND MEASUREMENT EXPLANATIONS

[PD] Validate original-customer outcomes through a channel independent of the disputed feed. If an operating trial follows, preserve unchanged eligibility, promises and remedies, and measure shared-resource spillover as well as the targeted outcome.

[HR/AI] The record does not identify historical intervention effects, validated root causes or the reporting lead's ownership of the underlying operating programmes.

EXHIBIT 5 / SCENARIO AND VALUE DISCIPLINE

A failed scenario rules out that scenario, not every alternative

[AI] A selected tail-to-target exercise is a response assumption. It becomes a feasibility bound only if it covers the entire admissible action set.

Bridge	Required evidence
Selected sites to network contribution	Original exposure weights, defect counts, selection rule and intervention reach.
Scenario response to feasible improvement	A stated response limit, prerequisites, capacity, interaction and spillover.
Rate reduction to avoided events	A common numerator, denominator, cohort and mature observation window.
Defect-pool value to economic benefit	Incremental causal prevention, avoidable unit cost, overlap and full implementation burden.
Automated report generation to savings	Measured effort and rework before and after, plus build, upkeep and displaced work.

[AI] A small share of sites can carry a large share of exposure or defects. A station-count allocation is not a measured volume-weighted rate. An intervention stronger than the simulated tail-to-target response could have a different result, while an apparently large arithmetic opportunity may be infeasible or expensive.

[HR/AI] The source's programme summaries and defect pools help frame discussion. They do not establish that the reporting process caused customer improvement or released cash. Prior-period actuals cannot fill an absent current-period result without a clearly marked temporal substitution.

KEEP THE BUSINESS CASE CONDITIONAL

[PD] Price only the mature incremental outcome under a credible comparison. Show realized cash, reusable capacity and decision-resource value separately. Never annualize a selected monthly pool or add an information benefit to the same avoided loss twice.

[AI/PD] No historical cost or response parameter calibrates the proposed model. A failed selected scenario remains informative without becoming a universal impossibility claim.

EXHIBIT 6 / CONTRIBUTION AND PROVENANCE

Credit the reporting design at the stage the evidence supports

[HR/AI] The supported contribution concerns structure, measurement disclosure and decision preparation. It does not transfer programme ownership or operating outcomes to the report author.

Evidence or role	Bounded interpretation
Registry description of specification and reporting work	Provisional registry-level hoop for assembling the mechanism and basis decisions.
Claimed document metadata and automation hooks	Leads for authentication; exact original artifacts remain unrecovered.
Goal choice or forum request	Authority is separate from authorship and source accuracy.
Operating playbook and programme actuals	Underlying owners retain execution and outcome responsibility.
Related planning artifact	Unresolved version; cannot authenticate monthly release, adoption or results.

[AI] A machine creator, last-modified field or successful file generation does not establish sole intellectual authorship, sending, acceptance or use. Likewise, a report's presence in a derivative register is not direct evidence that every monthly control ran. A bounded unsuccessful search is not proof that no later report existed.

[PD] Recover the versioned source, generated output, review exchange and release record. Match the decision to its as-of information. Record who selected the metric and target, who approved publication, who accepted an action, and who supplied the mature reference.

PRESERVE THE USEFUL CONTRIBUTION WITHOUT BORROWING IMPACT

[AI] The case supports an evidence-reconciliation and operating-design lesson. Personal savings, causal customer outcomes, portfolio ownership and authenticated deployment remain unestablished.

[HR/AI] Exact monthly reports, number spines, action-matrix specification, build hooks and decision originals were not recovered. Related material provides context rather than independent confirmation.

EXHIBIT 7 / THE RELEASE DECISION

Compare the reporting rule with a free qualified release before buying review

[PD] VDEC separates better information from a better decision rule. An evidence package can be useful while its proposed interpretation is poor.

Proposed control	Decision it governs
Original frame and version bridge	Which comparison is being made, under which goal and observation clock?
Resolution and independent reference	What is known, what remains unresolved, and how can the conclusion change?
Locked reporting rule and prices	Which error is costly, when is a qualified conclusion acceptable, and who sets the limit?
Strong feasible alternatives	Does additional review beat qualified release and the best use of the same evidence?
Separate action evaluation	If the report changes operations, what authorized trial measures mature customer and resource effects?

[PD] In the invented benchmark, projecting the resolved sample improves on an unconditional decrease but costs more and makes more wrong conclusive claims than a free qualified report. More tracing can worsen this specific rule: with six resolved IDs, its flat estimate would require a noninteger positive count. That is a discretization failure, not evidence that better information is intrinsically harmful.

[PD] A likelihood-aware use of the same sample has lower expected decision loss, but it still carries conclusive-error risk. The owner must approve an absolute tolerance and validate the assumed resolution law. Relative superiority to a weak headline is insufficient.

NEXT ACTION: VALIDATE THE BASIS AND THE DECISION RULE

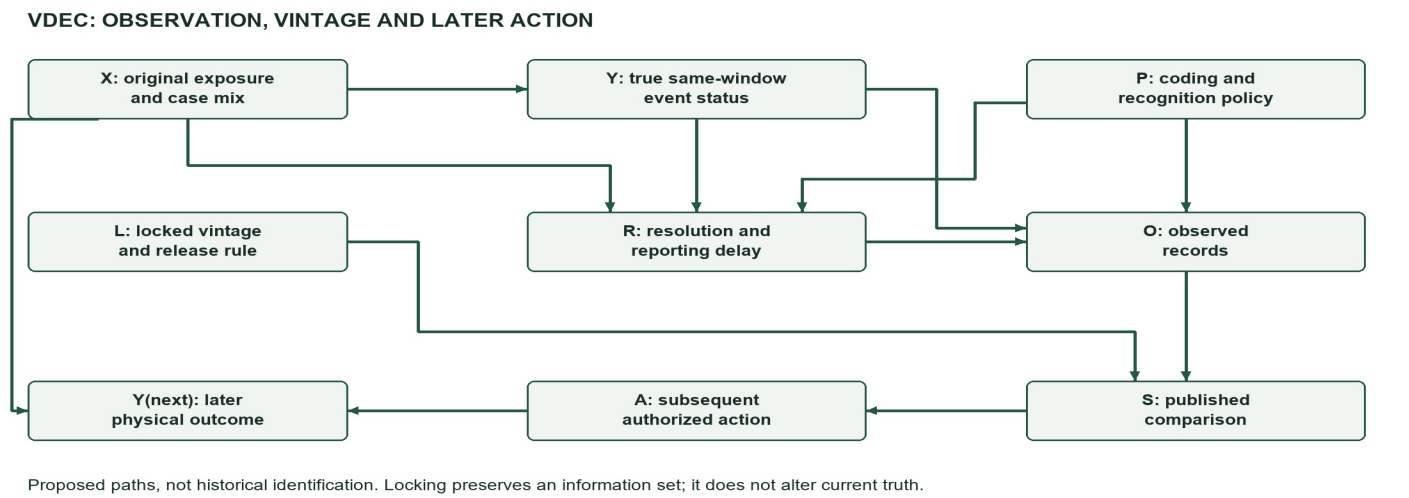
[PD] Begin with a bounded shadow release and independent adjudication. Keep the free qualified report as a confirmatory comparator. The modeled larger design addresses conditional precision; it does not establish recruitment feasibility or justify buying the tested projection.

[PD] Entirely hypothetical prices, priors, resolution and sample allocations. Reporting decision value is separate from physical improvement, historical time savings and Finance-recognized benefit.

TECHNICAL ANALYSIS

A reproducible vintage still needs a valid decision

[PD] The Vintage-to-Decision Evidence Contract (VDEC) binds original cohort identity, metric and policy version, ascertainment, decision-time information and the action justified by that information.



[PD] X denotes original exposure and case mix; Y is the true same-window event status; P is the coding or recognition policy; R is resolution and reporting delay; O is observed records. L locks a vintage, producing published S. S can affect decision A and later outcome Y(next). Locking changes neither current Y nor missing statuses. The diagram separates measurement paths from a prospective operational effect; conditioning on resolved records selects on R.

[HR/AI] The record describes a recurring monthly mechanism with number locking, decomposition, cause/action matrices and delivery artifacts. This parent case is distinct from a single monthly child. Exact mechanism originals were not authenticated in bounded recovery; a related planning artifact cannot establish their completed version, sending, adoption or execution. Byline and repeated narrative references remain one lineage. The contribution is a provisional registry-level hoop for framing and synthesis.

[AI/PD] Croushore and Stark motivate preserving the information actually available at a decision. Retrospective evaluation must replay that vintage, not substitute later corrections into earlier decisions. Preserve amendments alongside the original release, with the reason and affected conclusions. A source lock is a provenance control; accuracy requires independent identity, denominator, policy and outcome checks.

Basis: historical record [HR], analytical interpretation [AI], proposed graph [PD]. Croushore/Stark (1999); Office for National Statistics (2024). No historical numerical coordinates are public.

TECHNICAL ANALYSIS

The same locked decline admits opposite mature trends

[AI] Independent construction: two distinct original cohorts each contain 1,000 unique eligible units. Event status is binary. Resolved positives and negatives are correct; only listed unresolved identities can change status.

Hypothetical original cohort	A	B
Confirmed events	80	60
Unresolved original identities	20	80
Confirmed non-events	900	860
Mature event-count support	80 to 100	60 to 140

[AI] The locked observed difference is 60 minus 80 = -20. Let x_A and x_B be positive outcomes among the respective unresolved identities. The mature change equals $-20 + x_B - x_A$, where $0 \leq x_A \leq 20$ and $0 \leq x_B \leq 80$. Its sharp identified set is $[-40, +60]$ events. Every integer in that interval is attained by a feasible assignment; no probability model is required.

Same observed snapshot	Mature A	Mature B	Change
Stronger decrease	100	60	-40
No change	80	80	0
Reversal	80	140	+60

[AI] Correct status and exhaustive unresolved membership are essential. If prior A were truly closed at 80 and only current positives could arrive, the change would instead lie in $[-20, +60]$: the apparent improvement could only weaken. Retractions, duplicate records, unknown eligible identities or denominator changes require broader error bounds. Equal pull age does not ensure equal ascertainment. A caveat communicates uncertainty but does not resolve it.

[PD] Publish the original-cohort bound and the specific assumptions that narrow it. A decrease or increase is robust only when the entire admissible change lies on that side of zero; otherwise release a qualified comparison. This is descriptive trend identification, not the causal effect of a programme. Manski, Newman and Pepper motivate allowing an indeterminate performance verdict; the finite proof here is newly derived.

Basis: independently enumerated 1,701 status assignments [AI], wholly hypothetical. Sampling uncertainty and uncertainty about unobserved status are different.

TECHNICAL ANALYSIS

Validate the observation process before explaining the gap

[PD] Retain event time, source-record time, extraction time and amendment time. Freeze original eligibility and record which original identities remain unresolved at every vintage.

[PD] Names alone are not durable keys. Maintain stable identifiers and an effective-dated classification crosswalk; test rename, collision, split, merge and unmapped cases. Preserve uncertain matches instead of forcing closure. Independently verify exposure denominators, unit versus package grain, distinct events versus line items, cost basis and recognition policy. A discrepancy between a regional sum and its supposed whole requires reconciliation of scope and overlap before interpretation.

[AI/PD] Guenther and colleagues model changing reporting delays and evaluate forecasts against later references under explicit completeness assumptions. Here, build a cohort-by-vintage triangle before fitting a delay model; separate backfill, retractions, recoding and eligibility changes. Backtest out of time at actual release ages. A fixed lag or a label such as final does not establish closure. Model transport across workload, channel, site and policy versions needs calibration.

[PD] Draw a stratified reference sample from positive, negative, unresolved and amended original records, retaining inclusion probabilities. Two coders blinded to arm and report verdict assess event status, identity, eligibility and category; resolve disagreements using independent evidence. Report weighted confusion, prevalence, raw agreement and cluster-bootstrap kappa. Four hundred independent binary assessments yield worst-case simple-agreement precision near plus/minus 4.9 percentage points, not kappa power or clustered precision.

[AI/PD] For stratum weights w and rates r , a symmetric descriptive decomposition is $\text{sum}[\text{average}(w) \text{ times change}(r)] + \text{sum}[\text{average}(r) \text{ times change}(w)]$. It separates within-stratum and composition changes when the same partition and denominators hold. Das Gupta supplies the standardization logic; neither term identifies root cause. A missing weather label does not exclude an indirect weather pathway; ratios, tail ranks and fixed cause slots cannot replace linked mechanism evidence.

[HR/AI] A selected tail-to-target scenario tests only that action. Its shortfall does not establish that every stronger feasible tail intervention fails. Source scenario and regional-cost inconsistencies are retained as version or arithmetic questions. Retrospective policy subtraction and plan attainment do not independently identify causal physical savings or adherence.

Basis: proposed validation [PD] and source interpretation [HR/AI]. Guenther et al. (2021); Das Gupta (1993); Cohen (1960).

TECHNICAL ANALYSIS

Test reporting value separately from physical benefit

[PD] First evaluate a shadow reporting rule on all original monthly comparisons. A later operational trial asks a different question: whether decisions improve original-cohort service and resource use.

[PD] In the compact benchmark, the prior count is verified closed, while current original identities remain unresolved. Uniformly sampling identities does not make the resolved subset representative when positive and negative statuses resolve differently. Compare the projection rule with free qualified release, all fixed reports, robust observed-status bounds, the best rule using the same own-sample evidence and fuller tracing. Policies cannot see hidden status, evidence grade or other comparisons' signals.

[PD] Lock shadow reports at month close plus five days; reconstruct the same original window under the same policy at close plus 45 days, hidden from tuning. That reference horizon is a design assumption requiring field validation. The primary estimand is $E[L(\text{usual}) - L(\text{rule})]$ per original comparison, before subtracting fixed assigned review cost K in the same units. L prices mistaken directional reports and one-cycle qualification; report wrong conclusive claims on all assigned comparisons. Qualified release supplies bounds now, without withholding remedies.

[PD] Gneiting explains why scoring and decision objectives must align. The benchmark loss is an explicit judgment, not historical money or a claimed proper probability score. Charge unsuccessful tracing and unresolved cases. Separate review effort, avoided operating resource use and released cash. Freeze prices, rules and common-evidence rights before evaluation; test cheaper alternatives and price sensitivity even when they defeat the preferred design.

[PD] Randomize independent reporting teams or domains to approved shadow workflows, stratified by preassignment volume, data source, delay profile and policy version. Keep repeated comparisons within domain. Exact categorical moments and selected finite marginal tails assess the declared synthetic law; normal planning is conditional, not uniform coverage. Compare MDE and power at break-even with full-law clearance. Pilot variance, domain dependence, reference error and recruitable domains are needed before a confirmatory allocation.

[PD] Only a separately authorized workflow trial identifies $E[Y(\text{usual}) - Y(\text{action})]$ on original bookings. Preserve ordinary-flow harm, customer access and shared-team spillovers. Do not condition that total effect on postassignment resolution or adopted recommendations. A shadow reporting improvement alone demonstrates neither customer benefit nor programme causation.

Basis: prospective design [PD], no historical outcome estimate. Gneiting (2011); Hudgens/Halloran (2008). Model states and prices are independently hypothetical.

TECHNICAL ANALYSIS

Prespecify refutation, qualification and deployment limits

[PD] A contradicted claim, an unidentified claim and an underpowered comparison require different dispositions.

Proposition	Prespecified rejection or hold rule
P1: a locked trend identifies mature direction	Reject the general implication using compatible opposite-sign worlds; qualify field direction if its valid bound spans zero.
P2: equal pull age equalizes ascertainment	Reject when original-cohort reference reveals prespecified material differential capture; otherwise retain uncertainty about unobserved tails.
P3: a code decomposition identifies causes	Require independent linked mechanism and counterfactual evidence. Arithmetic closure alone holds causal attribution.
P4: a specified tail scenario excludes tail solutions	Reject with a feasible alternative exceeding its claimed achievable bound; one weak scenario cannot establish that bound.
P5: extra tracing merits purchase	Require full-cost gain and an agreed absolute report-error tolerance against strong cheaper alternatives. Imprecision holds funding; it is not zero effect.

[HR/RC] The private inventory contains 85 vectors: 23 source/sensitivity screens, 25 version comparisons, 19 design trajectories and 18 profiles. Source series have no fitted X or moving-range crossing and establish no stable phase. One ranked regional-cost profile has an X crossing, not a temporal operating signal. Tentative XmR uses 2.66 and 3.267. One mechanical two-network Laney sensitivity is ineligible; no eligible Laney panel is established. Missing observations are not fabricated.

[PD] Under an assumed linear conditional-mean model, expected untreated follow-up after selection at baseline x is $m + \rho \times (x - m)$, with population mean m and persistence ρ . This illustrates regression to the mean, not an estimate here. A later nonrandom rollout needs pretrend plots, equivalence bounds, negative controls and confounding/missingness sensitivity; nonsignificant pretrends do not prove comparability.

Limitations and identifying assumptions

[AI/PD] Exact originals, mature linked monthly panels and adoption evidence remain unavailable. Neither nowcasting nor standardization identifies an intervention effect without additional assumptions. Correct resolved statuses, exhaustive cohort membership, crosswalk validity, reference finality and stable loss preferences require validation. Reclassification, reporting access, selection and shared workloads can change conclusions. Claimed automation benefits require measured total work and error costs.

Basis: source/SPC scope [HR/RC], exact proof [AI], proposed falsification [PD]. Barnett et al. (2005); Manski/Newman/Pepper (2002).

PROPOSED EVALUATION / 1

Sample original unknown IDs and test the reporting decision

[PD] This synthetic shadow experiment changes a monthly conclusion, not physical events, customer promises or approval rights. No parameter is calibrated from the historical case.

[PD] The prior cohort has 10 verified, closed events. Current records contain 4 confirmed events and 8 unresolved original IDs. Hidden positive statuses H are 2, 6 or 8 with probabilities 0.40, 0.35 and 0.25: true totals are 6, 10 or 12. Three original comparisons share a hidden resolution grade, with probabilities 0.60/0.40, within each independent domain.

Context	IDs sampled	Positive / negative resolution by grade	Fee
Base	4	0.90/0.60 for both statuses	0.40
Selective-resolution stress	4	Positive 0.45/0.25; negative 0.90/0.60	0.40
Corroborated redesign	6	0.90/0.80 for both statuses	0.90
Fuller-review alternative	8	0.98/0.90 for both statuses	1.20

[PD] Sample uniformly without replacement. Policy sees positive count k, resolved count r and sampled count n, never H, grade or other comparisons' signals. Resolved labels are assumed correct. Usual always says decrease. The tested projection compares $4 + 8k/r$ with 10; $r = 0$ qualifies. The robust rule uses $[4 + k, 4 + k + 8 - r]$, retaining every unsampled and unresolved ID.

[PD] Wrong decrease/flat/increase costs 8/3/4; correct conclusions cost 0; qualification costs 1.5 for this cycle. All fees are paid on every assigned comparison, including no resolved sample. Qualification releases bounds; it does not withhold a remedy. Common intake and blinded research reference are separately budgeted.

LOCK THE REFERENCE BEFORE LOOKING AT RESULTS

[PD] Freeze the event frame at month close, the shadow report at close + 5 days and the hidden reference at close + 45 days. Reference reconstructs the same eligibility window and cannot tune the rule. Real finality and resolved-label validity require independent validation.

[PD] Invented decision-resource units per original comparison. The common and same-sample rules have equal evidence rights. Day 45 is a protocol horizon, not an assertion of universally complete truth.

PROPOSED EVALUATION / 2

A powered comparison with a weak headline does not justify purchase

[PD] The proposed projection is mixed with usual at probability λ per original comparison; the full review fee remains paid. This response path is a planning sensitivity, not observed compliance.

[PD] Randomize independent domains, each containing three comparisons, to usual or the fixed review rule. Exact grade-mixture moments preserve dependence. Directional gates require terminal-loss reduction above fee K and wrong-claim reduction above zero. Fixed normal radii are $1.96 \cdot \sqrt{V_{\max} / n}$, where n is domains per arm and $V_{\max}$ is the maximum two-arm variance over λ in $[0,1]$.

Context	Domains / arm	80% MDE	Break-even K	Zero-test power at K	Joint floor at full law
Base	16	2.215	0.40	7.2%	87.8%
Base	128	0.795	0.40	29.1%	>99.9%
Stress	128	0.793	0.40	29.2%	>99.9%
Corroborated	256	0.561	0.90	99.5%	>99.9%

[PD] MDE and K share terminal-loss units. The first three designs cannot detect the break-even effect with 80% power. The larger corroborated design can conditionally detect it, but recruitment of 512 domains / 1,536 comparisons is unestablished. Its primary rule still loses to free qualification. More observations cannot repair negative relative value.

[PD] Eighteen finite marginal tails use original-status likelihoods and exact finite probability polynomials, with numerical roundoff checks. Base/corroborated loss-null rejection is 2.5439%/2.5385%, slightly above nominal 2.5%. Economic-boundary false clearance is 2.5076%/2.3797%; positive zero-test detection there is 29.11%/99.44%. These are selected-law results, not uniform coverage or exact joint power.

MORE RESOLUTION NEEDS A BETTER RULE

[PD] With all six sampled IDs resolved, the projection can say flat only if $k = 4.5$, which is impossible; four resolved IDs permit $k = 3$. This arithmetic defect explains why more tracing can worsen this rule. It does not make improved information intrinsically harmful.

[PD] Normal-planning floors use a marginal union bound; zero is uninformative, not zero joint power. Eighteen finite tails use no Monte Carlo or sample-size search. Risk superiority to usual supplies no approved absolute error limit.

The free qualified report defeats the tested projection

[PD] Each entry is total priced loss including review fee / unconditional wrong-conclusive-claim risk. Lower is better. Every comparator uses the same state law within its context.

Reporting rule	Base	Selective stress	Corroborated
Usual decrease	4.8000 / 60.00%	4.8000 / 60.00%	4.8000 / 60.00%
Free qualified	1.5000 / 0.00%	1.5000 / 0.00%	1.5000 / 0.00%
Tested projection	2.0262 / 26.60%	2.6365 / 31.17%	2.7668 / 31.87%
Sampled robust interval	1.6229 / 0.00%	1.6229 / 0.00%	1.8408 / 0.00%
Best same-sample rule	1.2536 / 9.98%	1.3698 / 17.49%	1.1100 / 3.96%
Fuller review	1.2065 / 0.15%	1.2065 / 0.15%	1.2065 / 0.15%

[PD] All four fixed reports and every deterministic own-sample mapping are compared under locked prices. Priors and selective nonresolution likelihoods are assumptions, not a field oracle. Optimal mappings may accept greater error to lower priced loss. Free qualification retains zero conclusive-error risk. An owner-set absolute limit is still required.

[PD] Base projection gain over usual is 3.70680 favorable loss reduction minus 0.53304 added loss minus 0.40000 fee. Apply realization a and nonduplicated value share h only to favorable state-by-grade-by-category reductions: $net = a \cdot h \cdot 3.70680 - 0.53304 - 0.40000$. Every adverse category and fee stays fully charged. This is decision-resource sensitivity, not cash savings.

Realization a / net vs usual	h = 0.60	h = 0.80	h = 1.00
0.50	0.179	0.550	0.920
0.75	0.735	1.291	1.847
1.00	1.291	2.032	2.774

DO NOT HIDE THE STRONGER COUNTERFACTUAL

[PD] Even at full realization, projection loses 0.52624 to free qualification in base. At qualification price 0.50, free qualification also beats the paid optimal rule; at 3.00, the best free fixed report becomes flat. Reprice and reoptimize before choosing a workflow.

[PD] Within terminal-loss support 0.8 and declared truth states, reference reassignment on epsilon of original comparisons in each arm permits 16*epsilon loss and 2*epsilon risk-contrast error, with fees fixed. Wider frame errors require new bounds.

DECISION INFORMATION / 4

Separate the value of information from the quality of its proposed use

[PD] Better evidence has decision value only through a feasible rule. A poor projection is not an information-value optimum, and a model-derived optimum is not validated field practice.

[PD] At the main prices, perfect truth can reduce the best common-evidence terminal loss by at most 1.50000 per original comparison. The offered sample, used optimally, has pre-fee value 0.64638 / 0.53017 / 1.28998 in base / stress / corroborated contexts; after fees the gains over free qualification are 0.24638 / 0.13017 / 0.38998. These gains require the assumed prior, support and resolution law and the reported error trade-off.

[PD] Separately, choose a research study: validate capture, validate the metric/goal crosswalk, or defer. States are capture-study useful, crosswalk-study useful, or neither, with prior 0.40/0.35/0.25. Payoffs are (6,-2,-1), (-1,5,-2), and (0,0,0), respectively, in illustrative research utility. A signal names the true state with probability q and each other state with (1-q)/2.

Information experiment	q	EVSI	Study cost	Net EVSI
Low	0.50	0.46250	0.40000	0.06250
Base	0.75	1.45625	1.00000	0.45625
High	0.90	2.19250	2.00000	0.19250

[PD] All 27 signal-contingent research policies are enumerated. No-information utility is 1.45000 and EVPI is 2.70000. Low/base/high are different hypothetical experiments, not confidence intervals. Research utility is not added to the reporting loss reduction or to operating benefits.

VALIDATE BEFORE REGISTERING A CONFIRMATORY TEST

[PD] Enumerate original IDs; audit early resolved labels and late outcomes; measure selective nonresolution, domain dependence and costs. Freeze prices, error tolerance, sample rules and stronger comparators before evaluation. A later operating trial needs its own treatment, original-customer endpoint, interference unit and maturity design.

[PD] All information ceilings are conditional on this fixed support and action set. No historical causal effect, actual release decision, realized savings, achieved field power or approved risk tolerance is inferred.

REPRODUCIBILITY / READER PROTOCOL

Reproduce the release and the decision it permits

[AI] The Vintage-to-Decision Evidence Contract separates a frozen record from an identified performance conclusion. Each can be reproduced without treating the other as established.

Layer	Required record and independent check
Original measurement frame	Retain the original eligible unit IDs, event-window rule, policy, denominator, unresolved inventory and status corrections. Distinguish conceded units from unique customers and reported outcomes from physical service.
Release and revision lineage	Record source version, extraction time, reporting age, calculation and authorized target. Preserve earlier values. Classify new information, method changes and corrections separately; a FINAL label is not an independent adjudication.
Direction and decision boundary	Recalculate every credible matched-frame alternative. A conclusive direction requires all admissible completions to fall on the same side of the stated threshold. Otherwise release qualified bounds and assign the unresolved question.
Diagnostic and action link	Keep observed categories, causal hypotheses, intervention doses and delivery owners distinct. Recompute each scenario from its own baseline and exposure weights. A failed specified tail scenario does not exclude every feasible tail action.
Information and resource value	Recreate signal access, resolution failures, sampling, action rules and complete review costs. Compare the proposed rule with free qualified release, the same-evidence alternative and additional tracing. A favorable result against a weak comparator is insufficient.

[PD] Run paired reporting-policy replays on identical original cohorts. Freeze the loss function and capture assumptions before examining a hidden later reference; preserve all unresolved and sampled-negative records. Then vary selective resolution, review cost, qualification cost, target basis and reference error. Test the chosen policy prospectively on independent review domains.

[AI] The private archive retains source-cell locators, conflicting quantities, exact calculations, sampling code and internal review. The public examples are independently constructed. They do not disclose indexed historical coordinates or identify a causal benefit from reporting automation.

Protected source register and bounded recovery record; proposed VDEC reproduction protocol. Raw inputs, identities and source locations remain private.

REFERENCES / 1

Primary sources and their role in this case

[AI] These sources inform reporting vintages, partial identification, decision loss and evaluation. They do not authenticate internal releases or establish historical impact.

NIST/SEMATECH. Individuals Control Charts, e-Handbook section 6.3.2.2. Comparable ordered observations are necessary for operating limits. A revision vector or changing measurement phase is not a stable process series.

Laney, D. B. (2002). Improved Control Charts for Attributes. *Quality Engineering*, 14(4), 531–537. Use actual defect counts and exposure for rational subgroups. A two-network snapshot cannot establish temporal dispersion or process stability.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, 37–46. Blindly code the same original records under a fixed status rule. Agreement, ascertainment and construct validity are different properties; agreement cannot prove that missing records are harmless.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the Mean: What It Is and How to Deal with It. *International Journal of Epidemiology*, 34, 215–220. Selection of extreme regions can produce apparent recovery without a treatment effect. Fix the selected cohort and report reliability sensitivity when repeated controls are unavailable.

Callaway, B., & Sant’Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, 225, 200–230. For a later phased mechanism rollout, use coherent group-time effects on matched reporting and outcome clocks. Do not attribute a vintage change to the reporting mechanism.

Collier, D. (2011). Understanding Process Tracing. *PS: Political Science & Politics*, 44, 823–830. Attribute specification, report production and decision authority separately from target setting, operational execution and results. Editing metadata is not conclusive sole authorship.

Strong, M., Oakley, J. E., Brennan, A., & Breeze, P. (2015). Estimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample. *Medical Decision Making*, 35, 570–583. Value a validation study by the feasible decision it changes. Separate information value from operational impact and include a free qualified-release alternative.

Original author, journal, institutional and official methods sources. Read scope and case-specific application are retained privately.

REFERENCES / 2

Primary sources and their role in this case

[AI] These sources inform reporting vintages, partial identification, decision loss and evaluation. They do not authenticate internal releases or establish historical impact.

The W. Edwards Deming Institute. The Funnel Experiment. A routine movement is not a reason to retune operations. Establish comparable measurements and process phases before interpreting a directional change.

Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. A report can change recording behavior as well as real performance. Preserve original eligibility and independent outcome ascertainment when an indicator becomes consequential.

Goodhart, C. A. E. (1992). The Objectives for, and Conduct of, Monetary Policy in the 1990s. Reserve Bank of Australia Annual Conference, section 1(b)(i). Targets can alter the relationship between a reported measure and its underlying process. A favorable headline under one target basis is not proof of better operations.

Ittner, C. D., & Larcker, D. F. (2003). Coming Up Short on Nonfinancial Performance Measurement. Harvard Business Review, 81(11), 88–95, 139. Test the proposed link from input metrics through delivered action to mature outcomes. A detailed root-cause action matrix does not establish that causal chain.

Tucker, A. L., & Edmondson, A. C. (2003). Why Hospitals Don't Learn from Failures: Organizational and Psychological Dynamics That Inhibit System Change. California Management Review, 45(2), 55–72. Exceptions and missing data can support organizational learning when someone owns resolution. Closing an analytical task without checking the underlying failure can hide persistent problems.

OECD / European Commission JRC (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. Document denominators, weights and aggregation. Station count is not shipment exposure; separately normalized rates cannot be averaged into a network result without compatible weights.

Croushore, D., and Stark, T. (1999). A Real-Time Data Set for Macroeconomists. Federal Reserve Bank of Philadelphia Working Paper 99-4, June. Journal version: Journal of Econometrics 105 (2001), 111-130. Reconstruct the information available at each decision date and test conclusions across preserved vintages.

Original author, journal, institutional and official methods sources. Read scope and case-specific application are retained privately.

REFERENCES / 3

Primary sources and their role in this case

[AI] These sources inform reporting vintages, partial identification, decision loss and evaluation. They do not authenticate internal releases or establish historical impact.

Guenther, F., Bender, A., Katz, K., Kuechenhoff, H., and Hoehle, M. (2021). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal* 63(3), 490-502. doi:10.1002/bimj.202000112. Estimate time-varying ascertainment and backtest on preserved releases before using a maturity adjustment.

Manski, C. F., Newman, J., and Pepper, J. V. (2002). Using Performance Standards to Evaluate Social Programs with Incomplete Outcome Data: General Issues and Application to a Higher Education Block Grant Program. *Evaluation Review* 26(4), 355-381. doi:10.1177/0193841X02026004001. Permit an indeterminate performance verdict when credible outcome information does not identify which side of a threshold holds.

Das Gupta, P. (1993). Standardization and Decomposition of Rates: A User's Manual. U.S. Bureau of the Census, Current Population Reports P23-186. Washington, DC: U.S. Government Printing Office. Separate exposure-composition changes from within-stratum rate changes under a declared standard population.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 106(494), 746-762. doi:10.1198/jasa.2011.r10138. Freeze the report decision, loss and comparator before assessing whether additional evidence is worth buying.

Office for National Statistics (2024). The Office for National Statistics' Revisions Policy and Correction of Errors Policy, sections 3-4. Classify a changed figure by its reason: new information or methodology is distinct from correcting a mistake. Preserve the earlier release and explain the amendment. This is a governance analogy, not authentication of the source's disputed values.

Hudgens, M. G., & Halloran, M. E. (2008). Toward Causal Inference with Interference. *Journal of the American Statistical Association*, 103, 832-842. Original manuscript. Assign whole reporting teams or domains to contain shared data and reviewer spillovers.

Original author, journal, institutional and official methods sources. Read scope and case-specific application are retained privately.

FINAL REFEREE REPORT / INTERNAL REVIEW

Freeze the record. Keep the conclusion conditional.

[AI] Author-led, agent-assisted review; not external peer review. The case contributes a decision boundary for reporting on incomplete and conflicting operating evidence.

Major issue	Resolution and boundary
A locked value presented as truth	Treat version control as reproducibility. Preserve unresolved actuals and targets; release a direction only under an explicit eligible-cohort and ascertainment contract.
All changes called maturation	Separate additional events, revised scope, taxonomy changes, calculation errors and policy changes. The cause of an unexplained source disagreement remains unestablished.
A precise scenario with a mixed baseline	Reconcile the full scenario vector. A coherent alternative basis can explain rounding patterns without becoming an authenticated replacement; retain both sensitivities.
Codes or ratios treated as causal diagnoses	Use categories to select investigation. A coded location or small weather bucket neither identifies causal origin nor excludes an unmeasured path. Test interventions on mature outcomes.
More tracing assumed to improve the decision	Random selection does not eliminate selective resolution. Compare explicit rules with qualified release and charge every attempted review. Keep adverse cost and risk results.
Study precision confused with purchase value	Compare detectability with break-even under the same estimand. Even a precise comparison against usual reporting cannot justify a policy dominated by a cheaper feasible rule.

Partner critique: can a COO act on page one?

[AI] Yes. Appoint the release owner, freeze source and target versions, publish unresolved ranges and assign the next validation. Approve a reporting rule only if it improves the relevant decision at an acceptable error and resource cost. Operational action requires its own owner and evidence.

Remaining limits and minor issues

[AI] The technical layer consolidates authentication, capture, scope, attribution, selection, inference and reference limits. Only compiled descriptions of the unrecovered original artifacts are available. The finite replay demonstrates decision logic under stated assumptions; it does not estimate historical reporting accuracy, automation savings or customer impact.

REVIEW DISPOSITION / PASS WITHIN STATED SCOPE

[AI] Publish the evidence reconciliation, original-cohort counterexample and prospective reporting-policy test. Keep disputed historical effects unresolved and preserve the cheaper qualified-release comparator.

Final source, independent numerical, methods, partner and actual-page reviews required. Internal publication review is separate from external assurance.