
OPERATING CASE STUDY / DATA, ANALYTICS & AI

Leading Indicators for Performance Management

When a stronger analysis still needs validation

An anonymized project case study of a service network's leading-indicator audit: reconciling reporting clocks, model populations and metric definitions before using analytical findings to direct operating investment.

PROJECT CONTEXT / DISTRIBUTED SERVICE OPERATIONS

BLUF / BOTTOM LINE UP FRONT

Test the strongest candidate. Hold causal and financial credit.

RECALCULATED FINDING

The fitted sample is 48.84% of all panel rows and 62.43% of rows marked adequate. The register applies the smaller percentage to the wrong denominator; selection remains unresolved.

CONTRIBUTION IN THE SOURCE RECORD

Question-setting, alignment challenge and publication direction. Registry-attested role.

EVIDENCE STATUS

Evidence-reconciliation case. Historical impact and adoption remain unestablished.

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

EXECUTIVE ABSTRACT

Test the strongest candidate; repair the measurement contract before assigning impact

An evidence-reconciliation case on a distributed network's leading-indicator audit. The analysis challenged its own scorecard, but the available register does not identify program impact.

CONTEXT | A scorecard embedded an untested causal chain

[HR] Leadership asked whether operating inputs anticipated customer defects and whether a network governance program deserved credit for reported movement. An analytical rebuild changed outcome dating, data scope, network coverage and available regressors. The register reports different conclusions after the rebuild and records several publication restrictions.

OBJECTIVE | Separate a useful marker from a controllable causal lever

[PD] Decide which indicators merit prospective forecasting tests, which remain descriptive controls and what evidence would support an intervention claim. Require a consistent definition of the event, population, exposure, data-availability clock and comparison before approving an effect estimate.

APPROACH | Reconcile units, versions, populations and identification

[RC] Recalculate sample retention and the operating-unit translation of a reported rate ratio. [AI] Audit the bundled rebuild, median-delay interpretation, metric substitution and stated launch specification. [PD] Define a frozen-origin forecast test and a separate randomized operating-package evaluation with explicit economics.

FINDINGS | The corrected analysis still needs a validation gate

[RC] The reported fitted sample represents 48.84% of the full panel but 62.43% of the source-designated adequate subset; the register attaches the smaller percentage to the wrong denominator. [AI] The rebuild cannot isolate the clock correction because other inputs changed. Common launch-time indicators also duplicate full calendar-time indicators under the stated specification.

DECISION VALUE | Release a validation plan, not historical causal credit

[PD] The analytical owner should recover the run-of-record data definitions and design matrix, freeze an eligible cohort, and validate the strongest candidate using only information available at each forecast origin. The operating sponsor should require a mature-outcome trial capable of resolving finance's break-even before scaling on a causal claim.

IMPLICATIONS | Credit analytical governance within the evidence boundary

[HR] The register associates the principal with commissioning the question, challenging the initial alignment and directing publication corrections; agent workflows produced the estimation and prose. [AI] This supports registry-attested analytical governance, not independently authenticated authorship, adoption, customer improvement or savings.

S1: selected project register, M01–M18 and linked evidence. HR historical record; RC recalculated; AI analytical inference; PD proposed design. Identity, raw values and exact locators remain private.

EXHIBIT 1 / INDICATOR DECISIONS

Give each indicator the decision role its evidence can support

[AI] A reported association can justify further testing without establishing a causal lever or a deployable forecast.

Candidate	What the register supports	Decision now
Resolution marker	[HR] Favorable within-unit associations and a favorable cross-unit result in one network; the second network’s cross-unit result is inconclusive.	[PD] Prioritize time-respecting forecast validation and a defined operating-package test.
Throughput proxy	[HR] The modeled ratio differs from the scorecard pass-rate metric. Signs vary by lag and outcome.	[PD] Keep the scorecard metric untested by this model; do not transfer the proxy’s coefficient.
Dispatch-exception marker	[HR] Its strongest reported relationship with loss is contemporaneous and partly shared by definition.	[PD] Retain descriptive monitoring; require genuine forecast-origin separation for a leading claim.

[AI] The cross-unit and within-unit analyses reuse underlying observations. They examine different relationships but do not constitute independent data replication. A favorable result in one group and a nonsignificant result in another also does not establish a group difference.

[PD] Freeze the marker definition, feature availability, horizon, maturity rule and action before evaluation. Compare incremental predictive value with a simple baseline on later data. Test decision thresholds, calibration and loss-weighted performance, rather than selecting a marker because its smallest in-sample p-value is impressive.

OPERATING IMPLICATION

[AI] The register supports a prioritized research agenda. It does not prove that changing the marker will change customer outcomes, or that the existing governance weights are optimal.

[HR] S1:M03–M10, M14–M16. [AI/PD] Interpretation of reused evidence, proxy validity and out-of-time validation; Shmueli and Cawley & Talbot in references.

EXHIBIT 2 / VERSION RECONCILIATION

A bundled rebuild cannot identify the effect of correcting the clock

[HR] Outcome dating changed alongside data source, population coverage, network mapping and regressor availability.

RECONSTRUCT THE MISSING VERSION COMPARISONS

Proposed audit • fixed construct, outcome scope, exposure and estimator

Population rule	Earlier clock	Revised clock
Earlier population	Registered version also changes other inputs	Counterfactual rerun not available in this case
Revised population	Counterfactual rerun not available in this case	Registered version also changes other inputs

The diagonal comparison bundles corrections; it does not isolate the clock.

[AI] The source attributes a changed headline to corrected timing. That is a plausible explanation, but the observed comparison changes several components together. A larger or differently selected panel can change a coefficient even if its time alignment is unchanged. A newly available proxy can change what is testable at all.

[PD] Re-run old and new time definitions on both fixed populations with an identical feature dictionary, outcome scope, exposure and estimator. The within-population difference tests sensitivity to dating; the within-clock difference tests sensitivity to population. Report their interaction instead of choosing a single preferred explanation.

[PD] Where outcome definitions or regressors differ, first construct a common admissible feature/outcome contract. Evaluate remaining specification changes in additional branches. A population–clock table alone cannot separate a third changing component.

PUBLICATION RULE

[AI] State “the rebuild changed the result.” Attribute the change specifically to timing only after a controlled reconstruction supports that attribution.

[HR] S1:M01–M02, M06–M07; narrative Diagnostic_Story and Analysis_Performed. [AI/PD] Factorial reconstruction is proposed; no counterfactual rerun was available.

EXHIBIT 3 / ORIGINAL RECALCULATED FINDING

The fitted-sample share changes when the denominator is corrected

[RC] Rebase the full analytic panel to 100. The figure describes reported row retention, not station coverage or model validity.

ONE NUMERATOR • TWO DIFFERENT RETENTION QUESTIONS

Relative row-count index • full panel = 100 • no raw sample counts shown



Fitted / adequate = 62.43% • Fitted / full panel = 48.84%

“Adequate” is the source flag; its criteria and per-model inclusion remain unverified.

[RC] Fitted rows divided by the full panel give 48.84%; the same numerator divided by rows marked adequate gives 62.43%. The register mislabels the smaller share’s denominator. “Adequate” is a source flag with unrecovered criteria.

[AI] Correcting that statement improves the accounting; it does not eliminate selection risk. Fit retention concerns rows, while transportability concerns the target operating population. Repeated days from well-instrumented units can dominate a model even when a substantial share of sites has sparse or absent coverage.

Required sensitivity	Question it resolves
Fixed eligible cohort	Does the association persist when definitions and units remain constant across versions?
Inclusion audit	Do instrumented and excluded units differ in size, risk, workload or outcome recording?
Common support	Can the proposed operating change be evaluated within observed combinations of input and context?

SCOPE BOUNDARY

[AI] These counts are the register’s reported fitted-sample accounting. They do not establish that every coefficient in the model family uses the same observations.

[RC] S1:M01 and M16; private/raw_calculations.py. Full panel = 100; all displayed shares are relative to protected source counts. No separately indexed groups are pooled.

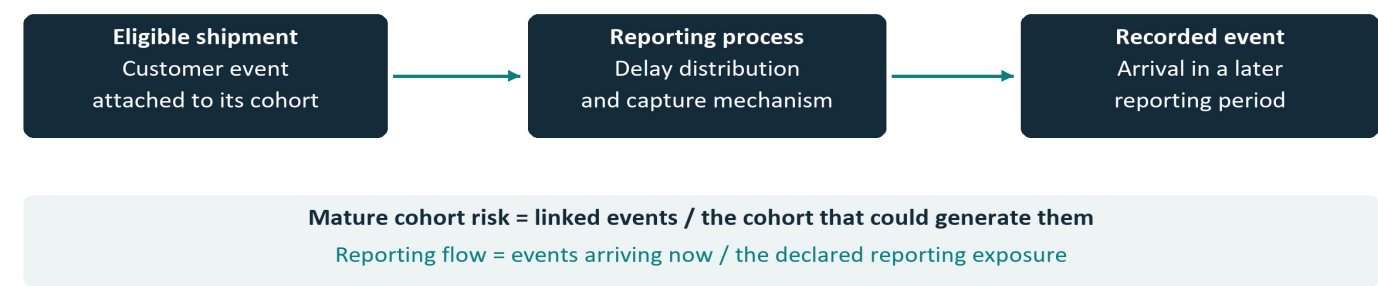
EXHIBIT 4 / OUTCOME TIMING

A reporting-date rate is not an eventual customer-risk rate

[AI] A median reporting delay helps describe the process. It does not certify that the most recent cohort is fully observed.

EVENT COHORT AND REPORT ARRIVAL ARE DIFFERENT AXES

Measurement contract • schematic, not an observed delay distribution



A median shift alone does not make these two quantities equivalent.

[HR] The source rebuild uses recorded-event creation dates and category-specific ship-to-recording delays. Its denominator is operating exposure dated by a separate activity clock. [AI] That combination can describe reporting flow per contemporaneous exposure; it does not automatically estimate eventual harm among the shipments in that period.

[AI] A delay median locates the center of the observed delay distribution. It is not a maximum delay, a guaranteed completion horizon or a correction for records not yet observed. The register’s claim that an observed rise is a conservative floor requires a matched-cohort accrual argument that is unavailable here.

Quantity	Correct comparison
Reporting flow	Records created in a period, with its declared exposure convention and scope.
Mature event risk	Customer events attached to a fixed eligible shipment cohort after a declared closure horizon.
Forecast performance	Predictions frozen at their origin, scored after outcome maturity using only then-available features.

REVIEW RULE

[PD] Store event time, record time and feature-availability time separately. Do not use a retrospective field that was unavailable when a supposed prediction would have been made.

[HR] S1:M02, M05, M11 and linked model descriptions. [AI/PD] Reporting-delay and forecasting interpretation; primary delay and prediction literature in references.

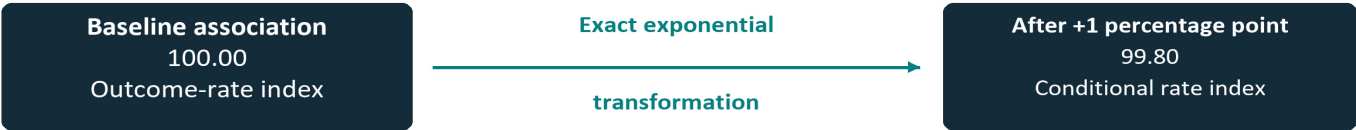
EXHIBIT 5 / EFFECT-UNIT DISCIPLINE

Translate the modeled increment before discussing operating materiality

[RC] The reported rate ratio is for a whole-unit change in a fractional predictor. A one-percentage-point change is one hundredth of that increment.

REBASE THE MODELED INCREMENT BEFORE TRANSLATING IT

Conditional modeled association • rounded source coefficient • no causal claim



Relative change = $\exp(\log(\text{reported rate ratio}) \times 0.01) - 1$
Recalculated: -0.197% | Source-reported: -0.198%

The higher-precision register coefficient reconciles the decimal; the original fit is unverified.

[RC] The rounded ratio implies a 0.197% lower modeled rate per percentage point. A higher-precision register coefficient reproduces 0.198%, reconciling the decimal without authenticating the original fit.

[AI] The calculation is a conditional association under the fitted model. It is not the expected effect of assigning labor, installing a tool or changing an operating procedure. A predictor’s apparent effect also cannot be transferred to a different scorecard metric merely because both were given the same label.

Translation gate	Required evidence
Construct and increment	Confirm the modeled field, scale and feasible within-unit change.
Population and support	Retain the fitted population and observed covariate support; flag extrapolation.
Outcome and exposure	Use a matched, mature customer-risk definition before translating to avoided events.
Value	Validate avoidable loss, incremental cost and overlapping credits before financial interpretation.

FINANCIAL BOUNDARY

[AI] Multiplying an unreconciled panel rate by a dashboard volume does not make an estimate portable. No historical savings are assigned to this case.

[RC] S1:M03, M15, E16; $\exp(\log(\text{IRR}) \times 0.01) - 1$ and $\exp(\text{beta} \times 0.01) - 1$. Public output is a relative modeled association; source coefficients and raw denominator conflicts remain private.

EXHIBIT 6 / IDENTIFICATION

Recover the design matrix before interpreting the launch coefficients

[AI] For a simultaneous launch, an event-week label is a relabeling of calendar week. The stated specification needs an identification check.

A COMMON LAUNCH DOES NOT CREATE NEW TIME VARIATION

Conditional rank audit • applies when calendar-week controls are saturated

Calendar-week indicator
1[week = launch week + k]

=

Common event-week indicator
1[event week = k]

Same column • same time contrast • no additional identifying variation
Recover the actual formula, dropped terms and comparison before interpretation.

An alternative trend model or an interaction would be a different specification.

[HR] The register describes common event-week indicators and station plus calendar-week controls. [AI] If those controls are saturated week fixed effects, adding common event-week indicators contributes no independent time contrast. Software must omit terms or use a different specification. The original design matrix was not available to resolve which occurred.

[AI] A second source ambiguity concerns coefficient units: a log-rate coefficient cannot be converted to an absolute defect-rate change by multiplying it by the reporting scale alone. The link function and baseline are required. These are specification questions, not evidence that the program failed.

[PD] Retrieve the estimation formula, omitted categories, rank, sample and output units before interpreting event-study magnitudes. Use the audit to define a prospective randomized comparison or a justified comparison-series design. A pre-existing movement weakens a simple before/after credit claim, but neither a favorable pre-period coefficient nor a nonsignificant post-period coefficient estimates the counterfactual.

DECISION

[AI] Keep program impact unestablished. Do not convert the unresolved specification into either causal credit or proof of no effect.

[HR] S1:M13 and linked source specification. [RC/AI] Common-time rank identity, demonstrated with a synthetic matrix in private/raw_calculations.py. The historical model was not re-estimated.

EXHIBIT 7 / DECISION CHARTER

Separate forecast validation from the test of an operating intervention

[PD] A useful warning signal and an effective action are different deliverables with different acceptance criteria.

Stage and owner	Evidence to release	Decision
Measurement / data owner	Versioned event, reporting, availability and exposure definitions; reconciled fit population; recovered model specification.	Authorize a reproducible analysis dataset.
Forecast / analytical owner	Frozen-origin predictions, later holdout periods, fair baseline, calibration, coverage and action-relevant forecast loss.	Use a candidate in a controlled decision trial if it adds prospective value.
Action / operating sponsor	A specified action package, compliance records and randomized rollout that contains shared-team spillovers.	Estimate the assignment effect on mature customer outcomes.
Economics / finance	Same-population effect, mature volume, avoidable loss, cost and credit reconciliation; a design that resolves break-even.	Scale only within the validated decision and operating boundary.

[PD] Maintain the distinction between observing a marker, receiving a warning and executing a response. A predictive improvement alone does not justify the cost of a prescribed action. An action can also help customers even if the marker that prompted it is poorly calibrated; the evaluation must identify which component is responsible.

[PD] Preserve rejected forecasts, missed opportunities, overrides and outcome revisions. Predefine safety and service guardrails, escalation authority and the minimum improvement that changes the investment decision.

COO ACTION

[PD] Commission the measurement reconstruction and a bounded validation sequence. Hold scorecard redesign and financial credit to the evidence required by each stage.

[PD] Proposed decision charter. S1 does not establish a completed out-of-time forecast test, randomized operating trial, scorecard adoption or finance-validated benefit.

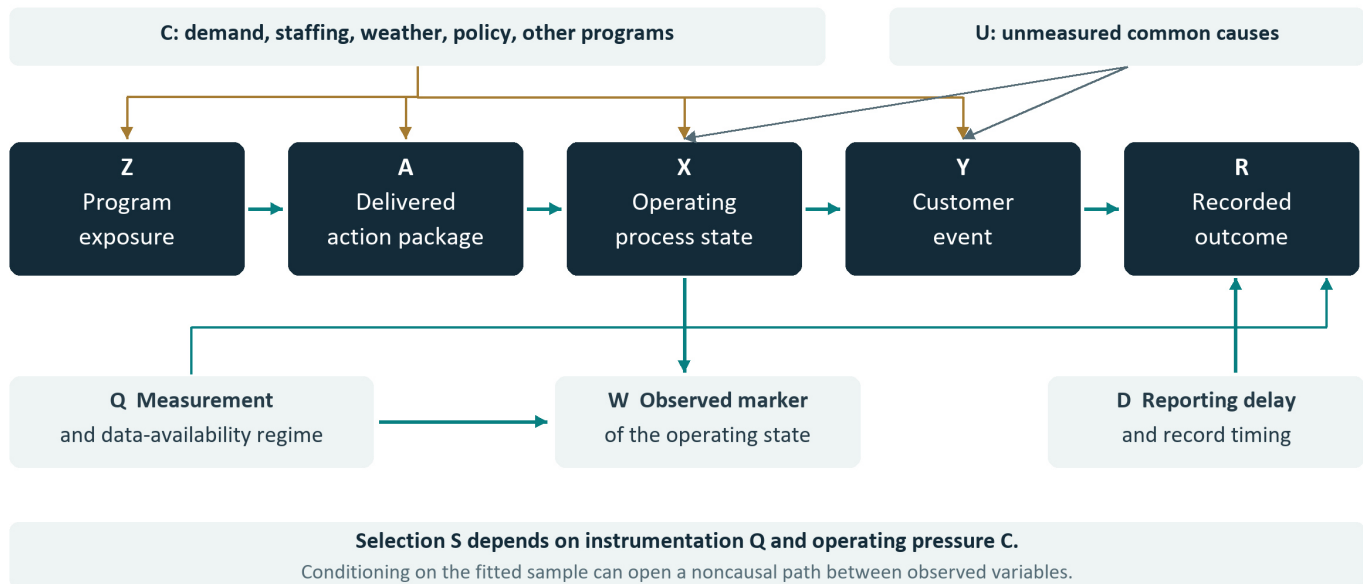
TECHNICAL FOUNDATION / THEORY OF CHANGE

The measured marker sits beside the action pathway

[PD] The reduced graph separates action delivery, operating state, customer events and the process that makes outcomes visible.

OPERATING EFFECTS, MARKERS AND RECORDED OUTCOMES

Proposed causal structure • solid arrows are hypotheses, not identified effects



[PD] Program exposure Z can change delivered action A, which can change operating state X and customer event Y. Marker W measures X; changing W in a dataset is not an intervention on X. Y and reporting delay D determine recorded outcome R. Measurement regime Q affects W and R.

[PD] Named common causes C can influence assignment, operating behavior, reporting and outcomes. Unmeasured U represents additional paths. Analytical inclusion S depends on instrumentation and operating pressure; restricting analysis to observed records can therefore change the associations being read.

[AI] The source coefficients concern W and R in selected records. They do not identify the delivered-package effect $A \rightarrow X \rightarrow Y$. The proposed trial randomizes package assignment, records delivered A and observes a defined horizon of Y. Unexplained residuals do not identify which omitted path generated them.

[PD] Proposed DAG based on S1's measurement and selection issues; Hernán & Robins and Shmueli in references. $C \rightarrow Q/D$ and the S subgraph are summarized in prose for readability.

TECHNICAL FOUNDATION · CAUSAL MODEL AND ESTIMANDS

A leading marker is not the treatment a manager can deliver

[PD] Separate the program, the action, the operating state, the observed marker and the recorded customer outcome.

[PD] Let Z be program exposure, A the delivered action package, X the operating process state, W its measured marker, Y the customer event, D reporting delay and R the recorded outcome. Hypothesized paths are $Z \rightarrow A \rightarrow X \rightarrow Y$, $X \rightarrow W$, and $Y, D \rightarrow R$. Measurement regime Q affects W and R. C includes demand, staffing, weather, policy and concurrent changes; U denotes unmeasured causes. S denotes analytical inclusion. Instrumentation and operating pressure can both determine S.

Quantity	Contract and interpretation
Descriptive movement	[PD] Change in recorded outcome rate for an explicitly fixed population, clock, denominator and window. Arithmetic alone does not identify an intervention effect.
Predictive value	[PD] Difference in prespecified forecast loss between a marker-enhanced model and a frozen benchmark on future eligible observations, using only information available at each forecast origin.
Action-package effect	[PD] $E[Y(a^*, H) - Y(a0, H) \mid \text{eligible}]$: a specified deliverable package versus current practice, measured on mature event cohorts over horizon H.
Program effect	[PD] $ITT(H) = E[Y(Z=1, H) - Y(Z=0, H) \mid \text{eligible}]$. Preserve assigned units, including noncompliers; do not substitute the effect of observed adherence.

[AI] Historical lag coefficients describe conditional associations on selected records. They do not identify the $X \rightarrow Y$ edge or either treatment effect. Conditioning on a post-treatment marker can block a total effect; conditioning on S can introduce selection bias. Residuals may represent unmeasured paths, error or chance, not a uniquely diagnosed cause.

Source S1: registered model and decision descriptions. Graph and prospective estimands are proposed; customer effects are not identified.

TECHNICAL FOUNDATION · CLOCK AND VERSION IDENTIFICATION

A joint rebuild cannot identify which correction changed the conclusion

[HR] The successor analysis changed the time alignment and analytical population; its regressor definitions also require reconciliation.

[PD] Define $\theta(P,C;M)$ as the same descriptive or model quantity under population P , clock C and fixed metric/specification M . Estimate old and revised clocks on both the common and expanded populations. Within each population, the clock contrast isolates a computational alignment change only when outcomes, regressors, exclusions, data vintage and estimator remain fixed.

Comparison	Interpretation and release gate
Two-by-two version audit	[PD] Populate all clock × population cells. The difference between clock contrasts reveals sensitivity to population; it is not a causal treatment interaction.
Regressor changed	[PD] Add a separate specification layer or restrict to a common metric. Do not attribute the joint diagonal contrast to time correction alone.
Event versus report	[PD] Link event cohort E , report date R and extraction vintage. Creation-date counts are an accounting flow; they are not automatically event-cohort risks.
Maturity and deployment	[PD] Estimate completeness $Fg(\text{age})$ by outcome and group. A median delay is a midpoint, not a mature cutoff. Freeze as-of availability before evaluating a forecast.

[AI] A report-date lag may reflect reporting delay or backfilled information rather than operational warning time. Censoring can bias estimates in either direction; no conservative-floor claim follows from the median alone.

[AI] The register describes common-launch event-week indicators alongside calendar-week controls. Saturated versions are collinear. Recover the design matrix, dropped columns and exact controls before interpreting coefficients; published labels do not resolve the specification.

Source S1: version chronology, maturity summaries and model descriptions. Factorial reconstruction and rank audit are proposed; missing comparison cells remain unestimated.

TECHNICAL FOUNDATION • MEASUREMENT VALIDITY

A valid calculation can still measure the wrong operating construct

[AI] Metric names do not establish construct equivalence, common populations or stable denominators.

Metric family	Validity and denominator contract
Resolution success	[PD] Define resolved cases, eligible resolution opportunities, carryovers and observation time. Test whether case difficulty, eligibility or missing coverage changes across nodes and periods.
Throughput ratio	[PD] Reconcile output and input events to the same cohort and flow boundary. A ratio exceeding unity may reflect carryover or timing; it is not a pass probability.
Upstream failure marker	[PD] Audit shared events, numerator subsets and exposure with recorded-loss outcomes. Part-whole association cannot independently validate a leading mechanism.
Customer outcome	[PD] Deduplicate events; identify controllable categories and overlap. Align eligible exposure to event or reporting time consistently. Contemporaneous shipped volume need not represent the cohort generating reports.
Composite and coverage	[PD] Inventory actual scored variables, normalization, weights, missingness and substitution rules. A model of a different variable does not validate the scored component.

[PD] For coefficient beta and feasible change delta, exact relative change is $\exp(\text{beta} \times \text{delta}) - 1$. A proportion, throughput-ratio unit and index point are distinct. Rebase coefficients when indexing. Reconcile population before translating relative associations into absolute outcomes.

[PD] Test taxonomy reliability with masked records sampled across categories, node types, vintages and disputes. Two coders independently apply a frozen handbook; adjudicate after scoring. Size the audit for a prespecified interval width. Report agreement, prevalence, confusion matrix and Cohen’s kappa with cluster-aware uncertainty. Agreement does not prove construct validity.

[AI] Low within-unit variation restricts identification and feasible extrapolation; it does not prove that an operational intervention is futile.

Source S1: regressor-definition defect, shared-event warning and unresolved population levels. Reliability and invariance protocols are proposed.

TECHNICAL FOUNDATION · SELECTION, MODERATION AND VARIATION

The fitted population defines where the association can be read

[AI] A larger assembled panel does not establish representative fitted coverage or wider causal transportability.

[PD] Reconcile assembled, eligibility-qualified, complete-case and specification-specific fitted rows separately. Track lag loss, unavailable markers, missing outcomes and excluded units. Compare retained and excluded units on pre-period levels, trends, volume and instrumentation. Weighting requires a defensible inclusion model and overlap; it cannot recover unobserved outcomes by assumption.

[PD] For bottom-ranked units, fix selection before follow-up and retain non-focus comparisons. A synthetic stationary Gaussian repeat-measure sensitivity illustrates the risk: selected excess $d = 10$ normalized points and reliability r imply expected reversion $(1-r)d$. Assumed $r = 0.40, 0.70$ and 0.90 give 6, 3 and 1 points. These invented inputs are not estimated project recovery.

[PD] Treat network type, node throughput, country, case mix and instrumentation as moderators. Test prespecified interactions and common support; different subgroup p-values do not establish different effects. Keep predecessor-version subgroup results quarantined until rerun under the successor clock and population.

[AI] The available before/after aggregates cover different windows and phases; event-study coefficients are model outputs, not repeated process measurements. They do not supply a commensurate series for natural-process classification. No historical movement is labeled an SPC signal here.

[PD] For future commensurate ordered observations, tentative XmR uses adjacent MR, limits $\text{mean}(x) \pm 3 \times \text{mean}(\text{MR})/1.128$ and moving-range upper limit $3.267 \times \text{mean}(\text{MR})$. Very short series are non-diagnostic. With event counts and eligible subgroup denominators, assess binomial dispersion before Laney p' . Never invent counts from rounded rates or narrate within-limit movement as established improvement.

Source S1: panel and fitted-sample summaries, predecessor subgroup outputs and aggregate comparisons. Selection sensitivity is synthetic; future SPC requires suitable observations.

TECHNICAL FOUNDATION • PERSONAL CONTRIBUTION

Credit the review decisions without inventing analytical authorship or adoption

[HR] The register attributes commissioning and methodological decisions to the principal and analytical production to a team/workflow.

Contribution claim	Process-tracing grade and boundary
Commissioned the audit	[HR] The register describes the principal setting the question and audience. [AI] Provisional registry-level hoop support; an authenticated request and recipient record would strengthen the specific commissioning claim.
Flagged timing and required rebuild	[HR] A session-associated memory record attributes the correction to the principal. [AI] Provisional hoop support for review involvement; the assistant-maintained account does not uniquely establish human origination or completed delivery.
Bounded the publication claim	[HR] The register records decisions to frame a causal chain as a hypothesis and withhold program credit. [AI] Provisional hoop support; recover versioned comments and accepted changes to establish the actual decision path.
Corrected final artifacts	[HR] Acceptance and revision records are described in the register. [AI] Provisional hoop support for the attributed correction role; output metadata alone does not identify the human editor.
Built, deployed or achieved impact	[AI] Personal code authorship, scorecard adoption and mature customer impact are unestablished. Team-produced artifacts do not prove sole ownership, field implementation or causal results.

[PD] A hoop test checks a necessary condition; passing is not sufficient. A smoking gun would directly document the particular decision or action and discriminate against plausible alternatives. Doubly decisive evidence must also exclude those alternatives. No such stronger grade is assigned from register excerpts alone.

[AI] Repeated career narratives and linked summaries may inherit one memory source. They are not independent corroboration. This case preserves the distinction between analytical stewardship, implementation ownership and outcome attribution.

Source S1: commissioning, method-review, memory, deliverable and acceptance-record descriptions. Original provenance and human authorship remain unauthenticated in this case.

TECHNICAL FOUNDATION • PROSPECTIVE EVALUATION

Validate prediction and intervention in separate prospective tests

[PD] A useful warning model and an effective operating action answer different questions and can fail independently.

Stage	Assignment, estimand and decision contract
Clock reconstruction	[PD] Recover event/report links and data vintages. Reproduce the clock × population comparison with common metrics. Resolve denominator and design-matrix discrepancies before ranking levers.
Prospective prediction	[PD] Freeze features at forecast origin. Use rolling held-out time periods and node holdouts, with maturity-complete labels. Compare a frozen benchmark on calibration, prespecified forecast loss and operational alert costs.
Action-package trial	[PD] Specify staffing, workflow, dose and fidelity. Pair eligible clusters and randomize immediate versus delayed package assignment at an operating-authority level that contains spillovers. Controls start after primary event cohorts close.
Assignment effect	[PD] Preserve assigned clusters, including noncompliers, for the mature-outcome ITT. The synthetic planning model targets an average-cluster effect under equal exposures and common baseline; heterogeneous weights require replanning.

[PD] Preregister horizon, reporting-completeness threshold, exclusions, primary outcome, multiplicity and economics. Simulate power using cluster count, allocation, baseline risk, serial correlation, within-cluster covariance, compliance and delay tails. Compare the MDE with break-even on the same outcome and horizon. A failing precision design must be enlarged, lengthened or narrowed before it can settle investment.

[PD] If randomization is infeasible, group-time DiD requires justified not-yet/never-treated comparisons, documented timing, overlap and anticipation checks. Inspect pre-trend equivalence and sensitivity; non-significance does not prove parallel trends. The paired-trial power calculation does not power this alternative.

[AI] Historical sample size does not establish achieved causal precision. Known-covariance synthetic planning scenarios elsewhere in this case are assumptions, not recovered project parameters.

Proposed design informed by source limitations and contemporary group-time treatment-effect methods. Forecast and treatment estimands remain separate.

TECHNICAL FOUNDATION • INTELLECTUAL CONTRIBUTION

The Clock–Population–Action Contract makes metric claims falsifiable

[PD] CPAC is a proposed discipline: fix the outcome clock, eligible population and executable action before interpreting or valuing a metric.

[AI] Ittner–Larcker–Randall motivate testing strategic measurement systems against performance. Shmueli distinguishes explanatory and predictive modeling; a lag coefficient does not satisfy both. Campbell warns that consequential indicators can change behavior. This case identifies measurement and validation gaps; it does not demonstrate gaming or an intervention effect.

Testable proposition	Falsifying observation or boundary
P1 • Clock validity	[PD] With population and metric fixed, correct event/report alignment should improve preregistered out-of-time calibration when reporting delay drives the mismatch. No improvement, or worsening, rejects that prediction for the tested setting.
P2 • Construct identity	[PD] Replacing a proxy with the actual scored construct should change its operating interpretation when cohort boundaries differ. Demonstrated equivalence across groups and vintages would reject the alleged construct gap.
P3 • Action validity	[PD] A feasible package that changes the marker should deliver its prespecified mature customer benefit. A randomized interval excluding that benefit rejects the package claim at the tested horizon, not every underlying mechanism.
P4 • Population validity	[PD] A candidate validated only in instrumented nodes will lose calibration or net decision value where instrumentation and case mix differ. Stable prospective performance under matched definitions would reject that predicted transport failure.

[AI] The extension is the joint-version audit: simultaneous clock, population and construct changes cannot identify one correction’s contribution. Factorial reconstruction makes that attribution testable. This is a reproducibility contribution, not a new causal estimator or demonstrated program effect.

Primary anchors: Ittner–Larcker–Randall (2003), Shmueli (2010) and Campbell (1976). CPAC and propositions are proposed, not empirically established.

TECHNICAL FOUNDATION • CONSOLIDATED LIMITATIONS

The evidence supports an analytical correction case, not a customer-impact claim

[AI] Each remaining limitation restricts a particular interpretation or decision.

Limitation	Consequence for the published claim
Source recovery	[AI] Original panels, code, authorship and acceptance sequence are unauthenticated here. Register arithmetic is reproducible; the historical model is not reproduced.
Joint version changes	[AI] Clock, population and specification changes are not isolated. The revised headline cannot be assigned to timing alone; conflicting source versions remain visible rather than averaged.
Identification	[AI] Reverse causation, time-varying confounding, endogenous package choice and shared shocks remain. A simultaneous rollout lacks the stated comparison; event-time/calendar-time rank must be verified.
Measurement	[AI] Scored metric and fitted proxy differ. Numerator overlap, exposure alignment, reporting tails and unvalidated taxonomy restrict coefficient interpretation.
Inference and selection	[AI] Multiplicity procedures differ across analyses, dependence may exceed the reported clustering, and nested outcomes are not independent confirmations. Same-panel corroboration shares selection and measurement errors.
Transport and consequence	[AI] Missing marker coverage and version-specific subgroup outputs constrain generalization. Scorecard adoption, feasible delivered dose, operational benefit and realized financial value are not established.
Attribution and privacy	[AI] Personal contribution is registry-attested and role-bounded. Public indexing conceals absolute scale; raw-data replication requires authorized access to the separate private annex and original artifacts.

[AI] Imprecision does not establish no effect; low variation does not establish futility. A reported lag does not establish deployable foresight. Pre-launch change does not identify the post-launch counterfactual. Universal rollout need not preclude future learning through credible external comparisons or prospective redesign.

[AI] Measurement and evaluation decisions remain actionable. Mature customer impact and booked savings remain unestablished.

Source S1 and independent methodological review of the registered claims. This is the consolidated boundary statement for the case.

TECHNICAL FOUNDATION • REPRODUCIBILITY AND DEFINITION LEDGER

A conclusion is reproducible only when its clock, sample and unit travel with it

[PD] Preserve the versioned evidence contract, not just the headline coefficient.

Ledger object	Required replication record
Source and version	[PD] Source hash, artifact stage, extraction locator, data vintage, claim class, supersession reason and unresolved conflict. Original operational evidence governs when available.
Population and time	[PD] Eligible opportunity, node inclusion, event date, report date, extraction date, exposure basis, maturity rule, missingness and fitting exclusions.
Metric and model	[PD] Exact numerator/denominator, regressor scale, feasible delta, fixed effects, offset, lags, dropped columns, convergence, clustering and multiplicity family.
Decision and attribution	[PD] Decision owner, adopted versus proposed status, evidentiary threshold, contribution hypothesis and process-tracing grade. Preserve alternatives that explain the same record.
Protected transformation	[PD] Index basis, transformed coefficient, allowable relative view and inversion checks. Store original values and identity mappings only in the private annex.

[PD] Replication sequence: freeze versions; reconcile panel-to-fit counts; reconstruct common definitions; rerun the two-by-two comparison; inspect design-matrix rank; recompute exact exponential effects; evaluate maturity and missingness; repeat inference with suitable dependence and multiplicity; rerun protected exhibits; inspect every final rendered page.

[PD] Sensitivity must retain the conclusion under each credible definition before it is called robust. If the conclusion changes, publish the unresolved fork. Keep predecessor subgroup outputs separate and label approximate arithmetic based on rounded coefficients. Do not infer raw coefficients, missing counts or fitted cells from narrative precision.

[AI] Current reproducibility is limited to calculations from the frozen register and explicitly synthetic scenarios. Reproducing the historical estimation requires original input data, executable environments and source authorization. The private annex records that distinction.

Reproducibility contract for this register-based case. Protected public calculations and private raw-input provenance are separate artifacts.

DECISION MODEL / VERSION ATTRIBUTION

Two endpoints cannot identify what the clock repair changed

[PD] Every numerical scenario on these four pages is synthetic. The following cells are examples, not reconstructed historical results.

[PD] Let P identify the population rule and C the clock rule; keep the metric definition, outcome, exposure and weights fixed. Both examples move the old-version index from 100 to 108.

Scenario	P0 C0	P0 C1	P1 C0	P1 C1
A	100	108	100	108
B	100	97	111	108

[RC] Average the two within-population clock contrasts. Scenario A assigns +8 points to clock and zero to population; B assigns -3 to clock and +11 to population. Identical endpoint changes therefore admit opposite clock conclusions.

[PD] This is a descriptive version bridge. A changed regressor requires its own harmonized contrast; coefficients for different constructs cannot be subtracted. Recover the missing cross-cells before attributing the historical reversal to one correction.

COHORT MATURITY IS A DIFFERENT QUESTION

[PD] For linked event cohorts only, mature risk ratio = observed ratio × pre-completeness / post-completeness. An observed ratio of 0.94 with pre-completeness 0.95 yields 1.191, 0.992 or 0.893 when post-completeness is 0.75, 0.90 or 1.00. These scenarios cannot correct a creation-date flow series without event linkage.

Synthetic models D1 and D2. No missing historical cross-cell, delay curve or event panel is estimated.

DECISION MODEL / CUSTOMER ECONOMICS

The trial must clear the cost of acting on the indicator

[PD] Price the randomized action package against the same population and endpoint: 42-day risk of at least one prespecified customer defect per eligible shipment.

[PD] Assume equal eligible exposure per cluster and a common untreated risk indexed to 100. Normalize that population's untreated customer-loss pool to 100 value units. $Net = 100 \times r \times a \times (1 - o) - K$: r is relative ITT risk reduction; a avoidability; o economic overlap; K incremental package cost.

Risk reduction	a = 60%	a = 80%	a = 100%
2%	-1.04	-0.72	-0.40
4%	-0.08	0.56	1.20
6%	0.88	1.84	2.80

[PD] The grid holds overlap at 20% and incremental cost at 2. The 100-unit pool is before overlap; o removes value already credited elsewhere. Set o to zero if the pool is already exclusive. These shares adjust costs, not an already incremental causal estimand.

[RC] Base avoidability of 80% requires a 3.125% relative reduction to break even. At overlap of 0%, 20% or 40%, the threshold becomes 2.50%, 3.125% or 4.167%. A favorable association supplies none of these treatment effects.

MATCH THE ESTIMAND

[PD] The proposed horizon is 42-day risk, not ultimate loss. Equal exposure and common baseline risk align trial and economic weights. Assume unchanged expected loss per affected shipment; revalue a cost-weighted outcome if severity mix changes. If weights change, recompute power.

Synthetic model D3; costs, treatment effects, exposure balance and outcome horizon are planning assumptions. No realized customer benefit is estimated.

DECISION MODEL / PROSPECTIVE POWER

Detecting an effect is weaker than establishing economic value

[PD] Use paired randomized delayed start: one cluster per pair receives the action package; its control crosses over only after the last primary cohort completes 42-day follow-up.

Clusters / pairs	Zero-null MDE	Economic power	Planning gate
48 / 24	4.15 points	49.30%	Fails
96 / 48	2.93 points	78.43%	Fails
120 / 60	2.62 points	86.67%	Conditional

[PD] Use 4 baseline and 8 post calendar periods; follow every eligible cohort for 42 days. Analyze control-minus-treated pair series with a pair intercept. Assume independent pairs and known covariance $\Sigma = 2\sigma^2[p11' + (1 - \rho)AR1(\phi)]$, with $\sigma = 8$ index points, $\rho = 0.25$ and $\phi = 0.35$. Identical shared calendar shocks cancel; treatment-specific shocks do not.

[PD] MDE uses a two-sided 5% zero-effect test and 80% power. Economic power instead requires the lower 95% bound to exceed 3.125 points when the true effect is an assumed 6 points. Gaussian GLS conditions on the proposed covariance and a constant post-period effect. If effects vary over time, prespecify economic aggregation weights and recompute power.

[RC] Ninety-six clusters have MDE below break-even but miss the economic power target. The 120-cluster design clears it conditionally. Raising σ to 10 defeats that design; serial-dependence sensitivity also requires evaluation before funding.

[PD] Preserve ITT assignment, common eligibility and independent outcome capture. Pilot covariance, unequal exposure, spillovers and missingness; then size and calibrate the final trial. A 42-day endpoint does not assert complete ascertainment of ultimate events.

Synthetic model D4; analytic GLS verified by 100,000 multivariate simulations. Conditional known-covariance planning power, not observed or guaranteed power.

DECISION MODEL / INFORMATION VALUE

An audit is valuable only when it changes the next investment

[PD] A separate three-state model chooses an experimental pilot, measurement repair, or hold. Its decision-option units are not customer savings.

[PD] Readiness states are experiment-ready, repairable measurement defect, and nonactionable; prior probabilities are 0.30, 0.40 and 0.30. Pilot payoffs by state are 24, -14, -45; repair payoffs are -4, 16, -6; hold pays zero. Before an audit, repair has the highest expected payoff: 3.40.

Signal accuracy	Study cost 1	Study cost 2.5	Study cost 4
Low: 50%	-1.00	-2.50	-4.00
Base: 75%	3.04	1.54	0.04
High: 90%	6.74	5.24	3.74

[PD] Each state generates its matching signal with the stated accuracy; errors split equally across the other two signals. Recompute the best action for every posterior. Table cells are EVSI minus study cost, not realized value.

[RC] EVSI is 0.00, 4.0375 and 7.7350 units. Perfect information is worth 10.20. The weak audit never changes the repair decision and destroys its full cost. Better signals can select pilot, repair or hold; their assumed accuracy needs validation.

KEEP LEDGERS SEPARATE

[PD] This values a future information choice under specified beliefs and payoffs. Do not add it to hypothetical trial benefits or claim it as the historical value of catching an error. Include delay and implementation costs before commissioning the audit.

Synthetic model D5; exact finite-state Bayesian enumeration with EVPI bounds. Priors, signal accuracy and decision payoffs are assumptions, not evidence from this project.

A corrected analytical claim is a deliverable; realized impact needs separate proof

[AI] The case’s contribution is to improve what a decision-maker may conclude from the available evidence.

Label	Meaning
[HR] Historical record	A statement or artifact is described in the supplied project register. The unavailable original was not independently authenticated.
[RC] Recalculated	Arithmetic from source quantities or an explicitly synthetic model. The input class and permitted interpretation are stated.
[AI] Analytical inference	A reasoned interpretation, specification constraint or evidence boundary; not an identified historical causal effect.
[PD] Proposed design	A future test, synthetic assumption, protocol or framework. It is not represented as deployed or observed.

[HR] S1 is the supplied project register. E and M references identify evidence and metric records within the selected project. The private ledger preserves worksheet cells, source versions and identity mappings. The source workbook was not modified; referenced originals were not found in the targeted available inventories.

[AI] The new findings concern a denominator error, a bundled version contrast and a conditional design-rank problem. They change the decision from accepting a corrected headline to validating the measurement and identification contract. They do not establish that the source analysis was wholly wrong or that the operating program had no effect.

[AI] Shewhart–Deming distinguish stable process behavior from reactions to noise; model coefficients are not control-chart observations. Campbell–Goodhart motivate rechecking a marker after incentives change. Tucker–Edmondson motivate treating a recurrent failure as a system-learning problem. OECD/JRC guidance makes construct and weighting validity explicit. The case uses these mechanisms as design questions; it does not claim to have experimentally confirmed them.

[PD] Public model outputs contain only protected relative quantities and explicitly synthetic scenarios. The separate private package contains raw values, source locators, alternative calculations, code and the internal review record. Public indexing does not make an unreconciled outcome denominator economically interchangeable.

Evidence classes apply throughout. Public statements remain narrower than several assertions in the source register; the private claim ledger records those adjudications.

REFERENCES / REPORTING CLOCKS AND PREDICTION

Reporting, prediction and causal explanation require different evidence

Primary methodological sources support interpretation and design. They are not evidence of this project's operational results.

- R1. McGough, S. F., Johansson, M. A., Lipsitch, M., & Menzies, N. A. (2020). Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLOS Computational Biology*. Event-by-delay reporting triangles and historical as-of validation.
- R2. Stoner, O., & Economou, T. (2020). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, 76, 789–798. Separating incident events from their reporting process.
- R3. Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25, 289–310. Author manuscript; descriptive, explanatory and predictive goals require distinct validation.
- R4. Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8, 431–440. Abstract-supported distinction between correlation and valid treatment-effect substitution; no clinical effect is transported.
- R5. Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *JRSS B*, 57, 289–300. Declare inferential families; the original guarantee has independence conditions.
- R6. Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *JMLR*, 11, 2079–2107. Separate selection and final performance evaluation.
- R7. Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13, 253–263. Forecast-loss comparisons with dependence-aware uncertainty.
- R8. Hernán, M. A., & Robins, J. M. Causal Inference: What If. Author-hosted text. Treatment definition, counterfactuals, causal graphs, selection and identification.

Linked original papers, author manuscripts and institutional methods. Source access and scope are documented in the private research memo.

REFERENCES / EVALUATION AND DECISION VALUE

Prospective design must resolve the investment threshold

Primary methodological sources support interpretation and design. They are not evidence of this project's operational results.

- R9. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*. Author manuscript on group-time effects and appropriate comparisons.
- R10. Roth, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*. Limits of pre-trend screening.
- R11. Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28, 182–191. Rollout structure and correlation must enter power calculations.
- R12. Hemming, K., et al. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350, h391. Calendar time, implementation and cluster-design considerations.
- R13. 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*. Value of improved decisions under uncertainty.
- R14. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. Independent classification and chance-corrected agreement.
- R15. Collier, D. (2011). Understanding Process Tracing. *PS: Political Science & Politics*. Hoop, smoking-gun and doubly-decisive evidence tests.
- R16. 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. Extreme-cohort selection and repeated observation.

Linked original papers, author manuscripts and institutional methods. Source access and scope are documented in the private research memo.

REFERENCES / PROCESS CONTROL, INCENTIVES AND LEARNING

Measurement discipline connects the scorecard to operating judgment

Primary methodological sources support interpretation and design. They are not evidence of this project's operational results.

- R17. NIST/SEMATECH. e-Handbook of Statistical Methods: Individuals Control Charts. Adjacent moving ranges and individuals limits.
- R18. Laney, D. B. (2002). Improved Control Charts for Attributes. *Quality Engineering*, 14, 531–537. Extra-binomial variation in attribute charts.
- R19. Wheeler, D. J. (2012). *Exact Answers to the Wrong Questions*. SPC Press. Process behavior and statistical interpretation.
- R20. The W. Edwards Deming Institute. *The Funnel Experiment*. Institutional explanation of tampering with a stable process.
- R21. Campbell, D. T. (1976). *Assessing the Impact of Planned Social Change*. Original occasional-paper record, ERIC. Indicators under evaluative pressure.
- R22. Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice*, 91–121. Publisher chapter record and abstract; relationships can change when indicators become controls.
- R23. Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial services firms. *Accounting, Organizations and Society*. Validate assumed relationships between measures and value.
- R24. 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*; author working paper. Operational workarounds and system learning.
- R25. OECD / European Commission JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Construct definition, normalization, weighting and sensitivity.

Linked original papers, author manuscripts and institutional methods. Source access and scope are documented in the private research memo.

FINAL REFEREE REPORT / AUTHOR-LED, AGENT-ASSISTED

The revised case supports analytical stewardship and a testable next decision

[AI] Internal review considered executive actionability, causal validity, measurement, arithmetic, confidentiality and reproducibility. It is not external journal peer review.

Major issue	Revision and final interpretation
Retention denominator	Recomputed fitted/full and fitted/adequate shares. Described the adequacy flag as source-defined and the fitted counts as a reported pool, not a universal model sample.
Timing attribution	Separated the bundled rebuild from an isolated clock effect. Required common definitions and missing comparison cells before assigning a correction’s contribution.
Maturity and forecast claims	Removed the conservative-floor inference from a median delay. Distinguished reporting flow, horizon-specific cohort risk and as-of predictive value.
Model identification	Kept common-event/calendar collinearity conditional on the stated saturated controls. Required the original formula and outcome units; withheld historical causal credit.
Economics and power	Used a synthetic delayed-start cluster design. Distinguished detection from clearing economic break-even and tested covariance sensitivity.

[AI] Minor issues resolved: source-qualified wording, rounded versus unrounded coefficient precision, marker/action separation, indexing scope and attribution grades. Registry-associated review decisions remain distinct from authenticated human authorship and implemented operating changes.

[AI] Referee disposition: suitable as an evidence-reconciliation and prospective-design case. It reproduces register arithmetic and identifies conditional specification constraints. It does not reproduce the original econometric fit, establish forecast performance or verify historical customer or financial impact.

[AI] Partner critique: a COO can act on page one—test the strongest candidate and withhold causal or financial credit. The first commission is a measurement reconstruction with an accountable owner; subsequent release depends on prospective decision value and a mature-outcome design capable of resolving the investment threshold.

[PD] Reopen stronger claims only after the original data/code, fixed population, feature vintages, treatment delivery, horizon-complete outcomes and adoption records are available.

Three parallel workstreams supported evidence audit, methods and editorial challenge. Final checks cover source arithmetic, synthetic models, public privacy and rendered pages.