# International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 20 No. 4 (2024)





ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 20, No.4, 2024
© 2024 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

*Research Paper*

# REPORTED OUTCOMES OF PREDICTIVE MAINTENANCE AND INDUSTRIAL AI DEPLOYMENTS

*Critical evidence synthesis and a meta-analytic research framework*

**Ashutosh Jaiswal**

**Abstract**

*Published predictive-maintenance outcomes are expressed through technical accuracy, equipment reliability, production performance and financial consequences. These measures describe different stages of an intervention and often use incompatible denominators. This paper develops a structured approach to evaluating and synthesising such evidence. A purposively selected evidence set comprises a US manufacturing survey reported by NIST and three published industrial case studies. The review preserves study design, operational boundary, observation period and source location, and recalculates selected reported changes. The audit identifies arithmetic and reporting inconsistencies that materially affect interpretation: percentage-point changes are sometimes presented as percentages, one reported mean-time-between-failures improvement differs from the change implied by its own values, and one steel case uses different downtime and cost bases across analytical sections. These findings motivate an estimand ledger linking every effect to a population, intervention, comparator, outcome, horizon and uncertainty measure. The paper specifies when rate ratios, mean differences and response ratios are meaningful and when a structured synthesis is preferable to pooling. It presents explicitly hypothetical meta-analytic calculations to demonstrate weighting, heterogeneity and sensitivity. The evidence supports study-specific reported outcomes and a programme for more comparable evaluation; it does not establish a pooled causal effect or a GCC-wide benefit rate.*

Keywords: predictive maintenance; industrial AI; evidence synthesis; meta-analysis; downtime; maintenance economics; reporting quality.

## 1. Introduction

A published improvement in predictive maintenance can refer to an algorithm, a machine, a production line or a financial account. A model can distinguish abnormal signals accurately without creating a completed maintenance intervention. A completed intervention can reduce an asset's downtime without changing line output. Additional output can remain unsold. The meaning of an outcome therefore depends on where it sits in the operating and economic sequence.

The research question is how reported outcomes can be compared without losing that meaning. A percentage alone is insufficient. It requires a numerator, denominator, baseline, period, unit of analysis and comparison design. The same numerical reduction can describe fewer events, shorter events, fewer lost production hours or lower estimated cost. Those effects cannot be combined merely because each is expressed as a percentage.

This paper examines a bounded set of published evidence and develops a framework for a broader synthesis. Its empirical component is a source-linked audit of selected reported values. Its methodological component specifies extraction, classification, effect construction and pooling decisions. The distinction allows the paper to make a concrete contribution without implying that a comprehensive database search or a pooled industry effect has already been completed.

The intended application is industrial decision-making and future research, including GCC process industries. Transfer to that setting requires a comparison of assets, operating conditions, intervention resources and economic boundaries. The paper retains the original geography and design of each source. It does not convert international case evidence into a regional causal estimate.

### 1.1 Contributions

The first contribution is an estimand ledger. An estimand is the precisely defined quantity a study seeks to estimate. The ledger records the population, intervention, comparator, outcome, time horizon and treatment of events such as shutdown, withdrawal or changing exposure. It makes explicit why two apparently similar savings claims can answer different questions.

The second contribution is a numerical consistency audit. It compares reported before-and-after values with the stated percentage change and checks whether the same measure retains its definition across a paper. The audit preserves the source's value and stores the recalculation separately. It does not silently correct the published record or assume that a discrepancy reveals which underlying value is wrong.

The third contribution is a pooling decision framework. It treats statistical compatibility as one requirement among several. A random-effects model can represent variation in effects within a coherent question. It cannot make downtime cost, downtime duration and prediction accuracy into the same outcome. Conceptual comparability must be established before choosing an estimator.

The fourth contribution is a set of worked analytical examples. These explicitly hypothetical examples show how weighting, clustering, outcome dependence and baseline selection affect a synthesis. Their role is explanatory. They are not presented as additional studies or empirical evidence about the performance of predictive maintenance.

### 1.2 Scope of the evidence review

The examined evidence set contains four study units. One is the US manufacturing survey described in NIST AMS 100-34 and a related journal publication by Thomas and Weiss [1,2]. The other units are the tube-filling-machine case by Natanael and Sutanto [3], the RPA and machine-learning maintenance case by Patrício and colleagues [4], and the hot-rolling run-out-table study by Yun and colleagues [5]. The two NIST reports are treated as related reports of one evidence family rather than independent replications.

The selection is purposive and question-led. It brings together a survey and industrial cases with numerical operational or economic outcomes that can be examined in accessible sources. It is not claimed to represent all predictive-maintenance studies. No PRISMA flow count or comprehensive search yield is invented for this bounded review.

The evidence audit was conducted using the identified source versions available on 2 October 2026. Full-text PDFs were examined for the three cases and the NIST report; the related NIST journal record supplied the reported comparison discussed here. Source page and table locations are retained in the appendices. Where a value is recalculated, the input values and equation are stated.

The remainder of the paper moves from definitions to the source audit, then to effect construction and synthesis design. The appendices provide extraction records, proposed review instruments and numerical demonstrations. They distinguish completed source examination from the procedures proposed for a future comprehensive systematic review.

## 2. Outcome architecture

An industrial AI intervention can be represented as a sequence from information to action and consequence. Sensors and records supply data. A model produces an output. People or automated rules interpret that output. A maintenance action occurs, or does not occur. The operating system then experiences a consequence that may affect production and financial results.

Each stage has a different outcome family. Technical discrimination concerns how model outputs relate to labelled conditions. Decision performance concerns whether a useful action is identified and completed in time. Reliability concerns the behaviour of equipment over exposure. Production performance concerns saleable output and operating constraints. Financial performance concerns incremental cost, margin or cash within a defined horizon.

A review should preserve these families. A high F1 score can support a claim about classification on a specified test set. It does not directly estimate the effect on annual production loss. An observed reduction in repair cost can support a financial description within its comparison, while leaving the contribution of the algorithm unidentified if several practices changed together.

The technical scope of predictive maintenance is discussed in the review by Zonta and colleagues [17], while Hoffmann and Lasch examine organisational barriers through qualitative case evidence [18]. These sources provide context rather than additional numerical deployment effects in the bounded audit.

The proposed architecture therefore records both the primary outcome and the stage it represents. This allows a synthesis to describe the completeness of the evidence chain. A study can contribute useful information at one stage without answering every question about the intervention's economic value.

**2.1 Prediction and decision performance**

Prediction metrics depend on labels, class prevalence, test design and the operating threshold. Accuracy can be dominated by the common class when failures are rare. Precision and recall describe different error trade-offs. F1 combines them under a specific mathematical definition. None of these metrics includes the full cost of acting on an alert.

Decision performance adds timing and consequence. An alert can be technically correct but arrive too late for a useful intervention. A false alert can consume inspection effort or create unnecessary downtime. The review should extract the available warning horizon, the action taken and the resources required, where those observations are reported.

The unit of evaluation is also important. Thousands of correlated sensor samples from one event do not provide thousands of independent maintenance opportunities. A model evaluated on randomly divided time samples may answer a different question from one tested on later events or previously unseen assets. The extraction record should preserve the split and the unit used for performance calculation.

A technical result can therefore be included in an evidence map while remaining outside a synthesis of operational effects. The review should explain that decision explicitly. Excluding a metric from a particular pool does not imply that the metric is useless; it means that its estimand belongs to another part of the evidence architecture.

**2.2 Reliability and production performance**

Reliability outcomes include event frequency, time between events, repair duration and downtime exposure. These measures are related but not identical. Fewer failures can coexist with longer repair time. Shorter repairs can reduce downtime even when event frequency is unchanged. A synthesis should retain the component measures before considering a combined effect.

Production outcomes introduce the process boundary. An asset can be unavailable while a standby unit preserves output. A line can lose output while several assets are repaired simultaneously. Summing equipment downtime without accounting for overlap can overstate the production interruption. The study should identify whether its reported hours belong to an asset, line or plant.

Overall equipment effectiveness combines availability, performance and quality under the study's definitions. A change in OEE can therefore arise through several channels. A review should extract the components where possible and distinguish an absolute percentage-point increase from a relative percentage increase. Both can be calculated, but they answer different questions.

For process-industry application, production recovery and inventory buffering may further modify the consequence. A downtime reduction does not automatically equal an increase in annual sales. The review records whether the study measured output, estimated avoided output loss or inferred a financial consequence from an assumed rate.

**2.3 Financial and environmental outcomes**

Financial outcomes require a cost boundary. Maintenance expense, repair cost, total operating cost, avoided production loss and programme return can overlap. A review should identify whether a reported saving is an observed expenditure difference, a modelled avoided loss or a projected benefit. These categories should remain distinct in synthesis.

The comparator determines the result. A programme compared with reactive maintenance may show a different effect from one compared with a well-executed preventive programme. The implementation cost may be included in one study and omitted in another. A common percentage reduction can therefore conceal different net-benefit questions.

Environmental outcomes also require a physical boundary. Energy intensity per tonne differs from total energy use. A plant can become more efficient per unit while increasing total consumption through higher output.

Emissions estimates depend on the included sources and factors. A maintenance-related environmental claim should retain those definitions rather than borrow the financial outcome's boundary.

The proposed extraction instrument stores both the reported label and the review's classification. This permits the reviewer to preserve the source while identifying what the number actually measures. If the available text does not resolve the boundary, the outcome remains ambiguous and is handled through narrative or sensitivity analysis rather than an assumed conversion.

## 3. Methods for the bounded evidence audit

The audit uses a fixed list of four study units selected for accessible numerical reporting across different designs. Each report is examined for intervention description, comparison, outcome definition, numerical inputs and uncertainty. Related publications are linked before counting independent evidence. The approach is a structured critical review of this selected set.

The extraction process preserves the exact numerical value and its source location. A second field stores any recalculated change. A discrepancy field records whether the issue concerns arithmetic, unit, denominator, time period or conflicting presentation. This separation avoids replacing the published claim with an analyst's preferred interpretation.

The audit does not infer missing standard deviations, event-level data or exposure durations. A before-and-after total can be recalculated as a ratio when its two values are known, but that does not supply a sampling variance or a causal comparison. The calculation's status is stated alongside the result.

The broader systematic-review protocol is developed later in the paper using PRISMA 2020, PRISMA-S and SWiM as reporting references [6–8]. These guidelines inform transparency. They do not certify the present bounded selection as exhaustive, nor do they establish the validity of an individual industrial study.

### 3.1 Source identity and study families

A study can appear as a technical report, conference paper, journal article and vendor summary. Counting each report as an independent study would give that underlying project disproportionate influence. The review therefore assigns both report identifiers and study-family identifiers.

The NIST report and related journal publication are linked in one family because they concern the same maintenance survey programme and overlapping economic questions [1,2]. The review uses each report for the specific information it supplies. It does not add their reported comparison percentages as separate independent observations in a pooled estimate.

For industrial cases, identity can be harder to establish when companies are anonymised. The reviewer should compare asset descriptions, geography, dates, authors, sample size and numerical outcomes. Similarity can justify a duplication flag, but it should not be treated as proof without sufficient evidence. An unresolved overlap can be explored through alternative inclusion scenarios.

The study-family register is preserved even when no pooling occurs. It helps readers understand the breadth of the evidence and prevents a large bibliography from being mistaken for a large number of independent deployments. The relevant count depends on the question: reports, plants, programmes, assets and independent comparisons are different quantities.

### 3.2 Numerical audit rules

For a before value B and an after value A, the relative reduction is $100(B-A)/B$ when B is positive. The absolute difference is $A-B$. When B and A are percentages, the absolute difference is expressed in percentage points. These rules are applied consistently in the audit.

A ratio is calculated only when the denominator is meaningful. A percentage reduction from zero is undefined. A ratio of negative profit values can be difficult to interpret and should not be used as though it were a cost reduction. Unit conversion precedes the calculation and is recorded explicitly.

The audit also checks whether row totals reproduce a reported aggregate. A small difference can arise from rounding. A larger difference can reflect omitted categories, a changed population or an error. The reviewer should identify the discrepancy and its possible implications without inventing the missing explanation.

When a paper reports incompatible versions, the review can retain both with source locations and define a sensitivity. It should not choose the most favourable result or assume that the abstract has priority over a detailed table. Source hierarchy within a paper is a review decision that requires justification, especially when the numerical conclusion changes.

## 4. Evidence from the manufacturing survey

Thomas and Weiss report economic comparisons across groups of US manufacturing respondents defined by reliance on maintenance strategies [1,2]. The related journal record reports 18.5% lower unplanned downtime for the more predictive-oriented group within the comparison between the two less-reactive groups. The NIST report describes 85 returned survey responses before exclusions. These values belong to the source's sampling and grouping framework; the returned count is not assumed to be the denominator for every outcome.

The review classifies this evidence as an observational group comparison. It can support a reported association between strategy reliance and outcomes within the examined sample. The comparison does not by itself identify the effect of assigning a predictive-maintenance programme to an otherwise identical establishment.

Several interpretation questions follow from that design. Firms can differ in asset condition, management capability, industrial activity and willingness to respond. Strategy reliance can itself reflect those characteristics. A reviewer should preserve the grouping rule and outcome denominator rather than translate the reported difference into a universal programme-saving percentage.

The survey nevertheless contributes an important measurement perspective: direct maintenance expenditure and consequences of inadequate maintenance are separate economic categories. The present framework adopts that distinction while retaining the survey's geographic and observational limits. Its numerical comparison is not pooled with the before-and-after case totals examined below.

### 4.1 Survey selection and denominator discipline

A returned response is not necessarily a complete analytical observation. Exclusions, unanswered questions and outcome-specific eligibility can change the denominator. The review should therefore extract the number of observations for each result where it is reported, rather than reuse a headline survey count throughout the synthesis.

Survey participation may also be related to the topic. Establishments with active maintenance programmes or a recent problem may be more likely to respond. Weighting can address some differences when a credible sampling frame and auxiliary information exist, but it does not automatically remove unmeasured selection. The review records the study's approach without supplying an unreported probability sample.

Group definitions can further affect interpretation. A comparison based on expenditure shares differs from a comparison based on a binary programme-adoption indicator. Expenditure shares can change because reactive costs rise after a failure. This creates a possible relationship between the grouping variable and the outcome that needs to be considered in causal interpretation.

For synthesis, the survey's contribution is therefore kept in an association stratum. A future review could compare similar survey estimates if populations, definitions and adjustment sets were sufficiently compatible. It should not treat a cross-sectional group contrast as the same estimand as a prospective controlled intervention effect.

## 5. Evidence from the tube-filling case

Natanael and Sutanto report a tube-filling-machine implementation with before-and-after operating measures [3]. Table 10 gives OEE values of 72.42% and 85.53%. The arithmetic difference is 13.11 percentage points, while the relative increase is approximately 18.10%. Table 11 reports total monitored downtime of 2,520 and 901 minutes, implying a reduction of 64.25%. Other passages use 62.38%, creating a reporting inconsistency that the review retains explicitly.

The detailed table also includes a downtime category that increases. This is useful because an aggregate improvement can coexist with adverse movement in an individual category. The review therefore preserves the category structure rather than summarising the entire case as uniformly favourable.

The study reports other operational changes alongside the implementation, including changes involving people, materials and methods. The review classifies the result as a before-and-after programme comparison. The reported totals do not isolate the effect of the predictive model from the broader implementation and concurrent changes.

The audit uses the tabulated totals for the reproducible arithmetic and records the conflicting narrative percentage separately. It does not assert which unreported underlying dataset would resolve the discrepancy. The case contributes evidence of a reported operational comparison and a concrete example of why source-linked extraction matters.

### 5.1 Percentage points and relative change

The distinction between percentage points and relative percentages has practical consequences. Moving from 72.42% to 85.53% changes the measure by 13.11 points. Dividing that difference by 72.42% gives a relative increase of about 18.10%. Reporting only an improvement of 13.11% would leave the intended calculation ambiguous.

The same issue occurs in accuracy, availability and quality rates. A change from 90% to 95% is five percentage points and about 5.56% relative to the original rate. The reduction in the complementary error rate is 50%. All three calculations can be correct, but they describe different quantities. A review must identify which quantity is being used.

This distinction also affects pooling. Percentage-point changes can be combined as absolute changes only when the underlying outcomes and scales are comparable. Relative response ratios answer another question and depend on baseline values. A reviewer should not convert between them without retaining the original baseline and understanding the metric.

The proposed estimand ledger therefore contains separate fields for original before and after values, absolute change, relative change and the source's stated change. The calculation can then be reproduced and the reporting ambiguity remains visible. This is a small data-design choice with a substantial effect on interpretability.

### 5.2 Aggregation within the case

The total downtime reduction is a weighted consequence of category-level changes, where the baseline minutes determine each category's contribution. An unweighted average of category percentage reductions would answer a different question and could be dominated by small categories with large relative changes.

The review should therefore retain category values when available and reconcile them to the total. If the sum differs from the reported total, the difference should be recorded. The reviewer may be able to identify rounding or omitted categories, but should not fill the gap without evidence. An aggregate can be used with a discrepancy flag when its interpretation remains clear enough for the relevant purpose.

In Table 11 of the tube-filling study, the displayed category rows sum to 2,395 minutes before and 901 after. That pair implies 1,494 fewer minutes and a 62.38% reduction. The printed before-total is 2,520, which instead implies 1,619 fewer minutes and 64.25%. Thus the two percentages can be reproduced from two different displayed baselines. The audit establishes this reconciliation discrepancy; it cannot determine which baseline is correct without the underlying records [3].

Dependence is another concern. Downtime categories from one machine share operating conditions, staff and observation periods. They are not independent studies. Entering every category as an independent effect in a conventional meta-analysis would give the case excessive weight and understate uncertainty.

A multilevel or robust-variance method may address some forms of dependence when enough independent studies and appropriate information exist. It cannot create independence where the underlying evidence is one machine. In a small evidence set, a transparent within-study summary and a narrative account of category variation may be the more informative approach.

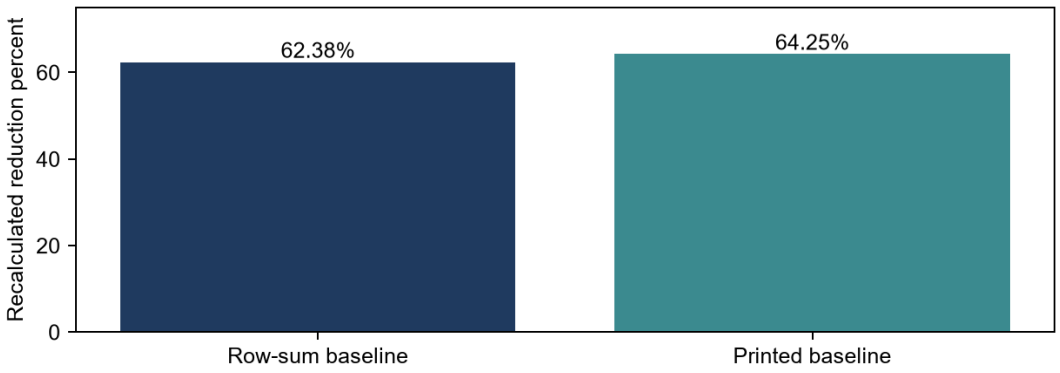



Figure 1. Two reproducible downtime reductions from the two displayed baseline totals in the tube-filling source. See Appendix C.

## 6. Evidence from the RPA and machine-learning case

Patrício, Varela and Silveira report a combined RPA and machine-learning maintenance programme [4]. Table 14 gives MTBF values of 23.47 and 70.4 hours per event, alongside a stated improvement of 100%. The ratio of the displayed values is approximately 3.00, implying an increase of about 200%. The review records that arithmetic discrepancy. The same source gives monthly maintenance costs of EUR 4,000 and EUR 2,500, a reproducible reduction of 37.5%.

The source describes company-provided before-and-after operating and cost information. The review retains that provenance. A published table can provide a clear numerical comparison while still leaving questions about measurement consistency, concurrent changes and causal attribution.

The MTBF discrepancy affects interpretation because doubling and tripling are materially different claims. The reviewer can verify the arithmetic of the printed values, but cannot determine from that calculation alone whether the percentage, the values or another part of the underlying record should be corrected.

The case is therefore included as a reported programme outcome with a numerical-consistency flag. Its maintenance-cost comparison belongs to a financial stratum. Its MTBF comparison belongs to a reliability stratum. Neither is treated as an independent replication of the other or combined with prediction accuracy into a single benefit estimate.

### 6.1 Event rates and time between failures

MTBF and event rate are reciprocal only under compatible definitions and exposure. If a study defines MTBF as total operating exposure divided by the number of failures, the reciprocal describes failures per unit exposure. A change in scheduled hours, excluded downtime or the definition of failure can alter the relationship.

The review should therefore extract the numerator and denominator behind MTBF where available. A monthly event count cannot be compared across periods without considering operating exposure. A plant that runs fewer hours may have fewer failures without a lower failure rate. Conversely, increased operating hours can coexist with improved reliability.

The mathematical transformation also changes the presentation of improvement. If the failure rate falls to one-third of its original value under fixed exposure, the corresponding time-between-failures measure triples. The relative rate reduction is about 66.67%, while the relative MTBF increase is 200%. Neither percentage should be substituted for the other without naming the metric.

For synthesis, a rate ratio may be more directly comparable across studies that report counts and exposure. Its variance requires appropriate event-count assumptions and treatment of clustering. The review should not infer a precise Poisson variance from rounded monthly averages when the underlying count and exposure history is unavailable.

### 6.2 Cost categories and combined interventions

The RPA case illustrates a broader problem in financial synthesis: several cost categories can move together and overlap. Repair cost per incident, maintenance cost per month and total operating cost per month have different denominators. Their percentage reductions cannot be added to produce a total programme saving.

The review's financial extraction asks whether the programme cost is included, whether the comparison period is the same, and whether an amount is recorded expenditure or an estimated avoided cost. It also asks whether the cost category includes labour, spares, software and external services. Missing detail is recorded as missing rather than replaced by a standard industry assumption.

A combined intervention should be described as implemented. RPA can change workflow and administrative effort, while machine learning changes information used in decisions. Training, process redesign and maintenance execution can also contribute. A before-and-after programme comparison measures the bundle under its observed conditions; it does not identify separate component effects without an additional design.

This is relevant to industrial procurement. A decision-maker comparing published results should compare the resources and operating model as well as the headline percentage. A low implementation cost in one setting may depend on existing infrastructure or internal expertise that another setting would need to purchase.

## 7. Evidence from the hot-rolling case

Yun, Choi and Lee examine an LSTM-autoencoder model for run-out-table equipment [5]. Table 18 reports downtime of 11.8 and 2.3 hours per year, implying an 80.51% reduction, and maintenance costs of USD

551,000 and USD 474,000. Elsewhere, Tables 16 and 17 use 2.3 hours as historical lost-production opportunity and a USD 500,000 maintenance-cost scenario. The review treats these as distinct reported analytical bases and flags the need for reconciliation before pooling.

The source contains both technical model evaluation and economic analysis. Those components provide different types of evidence. A classification metric on the study's evaluation data does not establish the same effect as an annual operating-cost comparison. The extraction keeps the two components separate.

The discrepancy is substantively important because the interpretation of 2.3 hours changes across the analytical presentation. If a value is a historical loss in one calculation and a post-intervention outcome in another, a reviewer needs a clear explanation of the populations, periods or scenarios involved before constructing a single effect estimate.

The paper therefore preserves the tabulated values and identifies their locations. It does not combine the alternative cost or downtime bases into a preferred estimate. The case contributes to the evidence map and to the methodological argument for a reconciliation stage before meta-analysis.

**7.1 Observed outcomes and modelled benefits**

An economic analysis can use observed failure history to model a future benefit. That is a legitimate analytical exercise when its assumptions are stated. It answers a different question from a measured post-implementation change. A review should classify the output according to how it was produced.

The classification should be made at outcome level. One paper may contain an observed technical test, a historical failure summary, an engineering estimate of avoided downtime and a projected expansion benefit. Assigning the entire paper a single label can obscure these differences. The estimand ledger therefore includes an evidence-type field for every outcome.

Modelled benefits can inform a synthesis of appraisal methods or assumptions. They should not enter a pool of observed intervention effects as though the model's predicted change were measured. If a review compares modelled estimates, it should examine shared assumptions and parameter sources, because common inputs can create apparent agreement without independent evidence.

The same principle applies to expansion scenarios. Multiplying a site result by the number of sites assumes transferability, comparable exposure and suitable scaling. A published expansion estimate should retain those assumptions. The review should not treat it as an additional deployment or a realised group-level outcome.

**7.2 Time horizons and operating exposure**

Annualised values are not necessarily annual observations. A short observation period can be scaled to a year, a historical average can cover several years, and a model can project a representative year. Each approach has a different evidential basis. The extraction should record both the observation window and the reporting horizon.

Rare failures make that distinction particularly important. A year with no event can occur by chance even when risk remains material. A short post-period may provide limited information on failure frequency. A before-and-after comparison should therefore retain event counts, exposure and uncertainty rather than rely solely on a percentage reduction.

The operating system can also change across periods. Product mix, speed, loading, environmental conditions and maintenance staffing can influence outcomes. A study may control or describe some of these factors. The review should identify what is reported and preserve unresolved differences as limitations of the comparison.

For GCC transfer, the relevant question is whether the receiving site shares the conditions needed for the effect mechanism. Geography alone is not the deciding factor. The review should compare process configuration, failure modes, response capability and economic context, while avoiding an unsupported claim that an international result predicts a local saving.

**8. Cross-study findings from the audit**

The examined sources report favourable outcomes under their respective definitions, but the comparisons occupy different evidential categories. The survey compares groups of establishments. The industrial cases compare periods or present operational and economic analyses within a deployment. Their numerical results cannot be treated as interchangeable estimates of one intervention effect.

The audit's first finding is therefore about comparability. Outcome labels conceal differences in unit, period and calculation. Downtime can mean minutes for selected components, annual hours for equipment or a

survey-based group measure. Financial outcomes can be per incident, per month or projected annual benefits. A synthesis must preserve those dimensions.

The second finding concerns internal consistency. The source-linked recalculations identify differences between printed values and stated percentage changes, and differences in the analytical bases used within a report. These issues do not establish that an intervention failed. They limit the precision with which a reviewer can state its magnitude and the confidence with which the value can enter a pool.

The third finding concerns attribution. The examined programme descriptions include operational processes and resources beyond an algorithm. Their results are most directly interpreted at the level of the observed programme and comparison. A claim about the isolated effect of industrial AI would require additional design or information.

**8.1 What can be compared directly**

The most direct comparisons within this evidence set are arithmetic comparisons within a source. A before-and-after pair with a clear unit can yield an absolute change and a relative ratio. That calculation is reproducible even when its causal interpretation remains limited. The audit therefore provides those calculations separately from an assessment of study design.

Across studies, a common label is insufficient. Two downtime outcomes can be compared descriptively when their boundaries are stated, but a pooled estimate requires a defensible common question. Asset-level monitored-component minutes and plant-level annual lost-production hours can differ in both scope and mechanism.

Financial measures are even more sensitive to context. Currency conversion cannot harmonise a cost measure that includes implementation in one study and excludes it in another. Inflation adjustment cannot resolve whether a value is an observed expense or a modelled opportunity cost. The review first harmonises the concept, then considers units and prices.

A useful evidence map can therefore present several outcome strata without an overall average. It can show where operational comparisons exist, where only technical performance is available and where economic consequences are modelled. This gives decision-makers a clearer picture of the evidence chain than a single pooled percentage spanning incompatible measures.

**8.2 Why no pooled causal effect is reported**

The bounded evidence set does not establish a collection of independent, conceptually aligned causal effect estimates with the uncertainty information needed for a defensible pooled causal result. This conclusion follows from the extracted designs and measures. It is a finding about the scope of the synthesis, not a claim that meta-analysis is inherently unsuitable for industrial maintenance.

The Cochrane Handbook describes requirements for effect estimation, synthesis and the handling of missing summary data [9–11]. Its methods are developed for intervention reviews, and their application here requires industrially appropriate outcome definitions. The statistical principle remains relevant: a weight cannot be derived responsibly from a sampling variance that has not been observed or justified.

An unweighted average of reported percentage improvements would not solve this problem. It would combine different outcomes and grant equal influence to studies regardless of their estimand, design and dependence. A very large saving from a narrow component case could dominate the narrative without being comparable to a broader survey contrast.

The paper instead provides a structured synthesis and a meta-analytic framework. A future comprehensive review could populate that framework with compatible estimates. Where a coherent subset exists, pooling can be considered. Where it does not, the synthesis should explain the grouping and reporting method, consistent with SWiM's emphasis on transparent alternatives [8].

**9. The estimand ledger**

The estimand ledger is the proposed core data structure. Each row represents one outcome comparison, linked to a study family and report. It identifies the population, intervention, comparator, outcome, period, unit, effect measure and handling of changes in exposure. It also records whether the result is observed, estimated by the source or recalculated by the reviewer.

This structure resolves several recurring ambiguities. A study can contribute multiple outcomes without those outcomes becoming independent studies. A modelled financial benefit can be retained without entering an

observed-outcome pool. A corrected arithmetic calculation can coexist with the source's printed value. Each decision is visible at the outcome level.

The ledger also records the target interpretation. A within-machine before-and-after comparison answers a different question from a between-firm adjusted association. Both may be relevant to an evidence map, but they belong to different design strata. A pooled result should state the design and population represented by its contributing rows.

Finally, the ledger connects effect size to uncertainty. It stores the standard error, variance, confidence interval or raw inputs used to derive uncertainty, together with assumptions about pairing and clustering. If those inputs are missing, the row remains in the descriptive evidence set and its eligibility for pooling is considered explicitly.

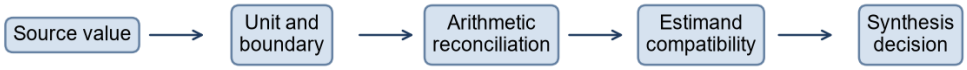



Figure 2. Proposed sequence from extraction to a synthesis decision.

**9.1 Population intervention comparator and horizon**

The population field identifies the actual assets, plants or establishments observed. It should include the eligibility criteria and any restriction to selected failure modes. A result from monitored components should not be relabelled as a whole-plant result. The population at risk can also change during the study, which affects the denominator.

The intervention field describes the implemented bundle: sensors, analytical model, thresholds, review process, maintenance response, training and support. The comparator field describes what was used before or in the comparison group. Reactive, preventive, condition-based and mixed practices should be recorded as the source defines them.

The horizon field distinguishes data collection, intervention exposure and outcome reporting. A six-month observation can be annualised, but the annualised value retains the shorter observation basis. A programme may be installed during part of the window and fully operational only later. These dates should be retained rather than compressed into a single publication year.

The ledger also specifies how shutdowns, retirements and missing operating periods are handled. Removing failed assets from the denominator after failure can make later performance appear better. Including periods when an asset was not scheduled to operate can alter a rate. A coherent estimand requires an explicit rule for these events.

**9.2 Outcome hierarchy and selection rules**

A study may report several versions of the same outcome: total downtime, selected-component downtime, unplanned downtime and downtime cost. The review should pre-specify which version is primary for each synthesis question. The choice should be based on conceptual fit and data quality rather than effect magnitude.

When several time points are reported, the selection rule should identify the relevant horizon. A short-term effect may reflect commissioning or intensive support. A later effect may reflect routine operation. The review can present both as a trajectory, with dependence recognised, rather than choose whichever is larger.

Adjusted and unadjusted estimates also require a rule. An adjusted estimate can address confounding, but different adjustment sets can target different effects. The review should retain the variables used and consider whether they include mediators. A more heavily adjusted model is not automatically the most appropriate estimate for every question.

The ledger makes these choices reproducible. It records the selected estimate, alternatives and the reason for selection. Sensitivity analysis can then examine whether the conclusion changes under a defensible alternative rule. This is preferable to silently treating the most prominent number in an abstract as the authoritative effect.

**10. Effect measures for industrial outcomes**

The effect measure should follow the outcome and design. A mean difference is suitable for a common continuous scale when group means and uncertainty are available. A rate ratio can compare events per unit exposure. A response ratio can describe proportional change in a positive continuous quantity. Each has assumptions and limitations.

For downtime exposure, a rate can be defined as qualifying downtime divided by scheduled exposure. The ratio of post- to pre-intervention rates then separates a change in downtime from a change in operating hours. If only raw downtime totals are available, the ratio of totals should be labelled accordingly and should not be assumed to be exposure-adjusted.

For cost outcomes, a response ratio can be meaningful when the cost boundary and price basis are comparable. Costs can be skewed, include zeros or contain rare catastrophic events. The transformation and uncertainty method should reflect those features. A logarithmic ratio is undefined at zero and should not be repaired by an arbitrary constant without justification.

For OEE and other percentages, an absolute change in percentage points can be clear, but it remains specific to the underlying definition. The review should avoid pooling a percentage-point change in OEE with a relative reduction in downtime. Standardising a scale does not erase a difference in construct.

**10.1 Log response ratios**

For positive before and after quantities B and A, define the descriptive log response ratio $y = \ln(A/B)$. Exponentiating y returns the ratio A/B. A ratio below one indicates a reduction when lower values are favourable. The corresponding percentage reduction is $100[1-\exp(y)]$. The direction of benefit must be defined separately for outcomes such as availability, where higher values are favourable.

This transformation is useful because multiplicative changes become additive on the log scale. It does not establish independence or causal identification. A before-and-after ratio remains a before-and-after comparison unless supported by a design that accounts for other changes.

The variance of a log response ratio depends on the data structure. Independent group means require information on sample sizes and variability. Paired observations require covariance information. A ratio of annual totals from one plant is not automatically a sample mean with a known standard error. The ledger therefore retains the raw basis of the ratio.

When a source provides only a percentage change, the reviewer can sometimes reconstruct the ratio but not its uncertainty. Such a result can enter a descriptive display. It should not receive an inverse-variance weight generated from an invented sample size or an assumed standard deviation presented as observed.

**10.2 Event-rate ratios**

Suppose a study records event counts $N_0$ and $N_1$ over exposure $E_0$ and $E_1$. The event rates are $N_0/E_0$ and $N_1/E_1$. The rate ratio is $(N_1/E_1)/(N_0/E_0)$. This definition makes exposure explicit and allows the direction and scale of the comparison to be inspected.

Under a simplified independent Poisson model with positive counts, the variance of the log rate ratio is approximated by $1/N_1 + 1/N_0$. That expression is a model-based result. Industrial events can be clustered, recurrent, dependent or overdispersed, so the assumption must be examined rather than adopted automatically.

The unit of exposure should match the event. Asset operating hours can be appropriate for asset failures, while production-line hours may be appropriate for line interruptions. Counting several dependent component failures as independent events can misstate both the rate and its uncertainty.

Zero counts require a declared method. A zero post-period count can be informative, but its interpretation depends on exposure and the event process. Continuity corrections, likelihood methods and exact approaches answer different statistical questions. A review should pre-specify the method and examine sensitivity where zero-event studies materially influence the result.

**10.3 Continuous outcomes and paired changes**

For a continuous outcome measured before and after on the same units, the variance of the change depends on the correlation between measurements. If standard deviations are $s_0$ and $s_1$ and the correlation is r, the

variance of an individual change is s_0 squared plus s_1 squared minus 2rs_0s_1. The variance of the mean change additionally depends on the number of independent units.

Ignoring pairing can overstate or understate precision depending on the data and the approximation used. Treating repeated sensor readings as independent units can severely overstate the effective sample size. The review should identify whether the independent unit is an asset, event, shift, plant or another level.

When the pre-post correlation is unreported, a sensitivity analysis can show the consequences of plausible assumed values, clearly labelled as assumptions. The primary analysis should not conceal that dependence. The Cochrane guidance on effect measures and missing summary data provides a methodological reference for this issue [9,10].

A standardised mean difference can facilitate comparison of different scales measuring the same construct. It does not justify combining unrelated constructs. Standardising downtime, prediction accuracy and operating profit would produce numerically similar units while leaving their substantive meanings incompatible.

**11. Meta-analytic models and heterogeneity**

A common-effect model assumes that the included estimates target one common underlying effect, with observed differences attributed to sampling error under the model. A random-effects model allows a distribution of effects across studies. The choice should follow the question and the evidence structure, not a mechanical test result.

For study estimate y_i with sampling variance v_i, a conventional inverse-variance common-effect weight is 1/v_i. A random-effects weight is $1/(v_i+\tau^2)$, where $\tau^2$ represents between-study variance under the model. Estimation of $\tau^2$ and interval construction require particular care with few studies. DerSimonian and Laird provide a foundational random-effects formulation [12]; later software and methodological developments offer additional estimators and diagnostics [13].

Heterogeneity can arise from differences in assets, programme implementation, comparators, outcome definitions and bias. A statistical variance term does not explain those differences. The review should examine the substantive grouping before fitting a model and retain that grouping in interpretation.

The pooled mean, when justified, describes the included effect distribution under the model. It is not automatically the effect expected at every receiving plant. A prediction interval can help express between-setting variation, but its reliability depends on the number and comparability of studies and the assumptions used to estimate that variation.

**11.1 Interpreting inconsistency**

The I-squared statistic is commonly used to describe the proportion of observed variation associated with heterogeneity under a meta-analytic framework [14]. It should not be treated as a universal quality score. Its value depends on study precision and the evidence set, and uncertainty can be substantial when few studies contribute.

A low inconsistency statistic does not establish conceptual compatibility. Several studies can estimate different outcomes with large standard errors and show little detectable statistical heterogeneity. Conversely, highly precise studies can detect small differences that may be operationally unimportant. The review should interpret the statistic alongside effect magnitudes and study characteristics.

The between-study variance is on the scale of the chosen effect measure. Its practical meaning should be explained through transformed effects or prediction ranges where appropriate. Reporting a variance parameter without a substantive interpretation gives decision-makers little guidance.

The bounded evidence set in this paper is not used to estimate I-squared or a pooled between-study variance. Its outcome and design differences are addressed first through stratification and the estimand ledger. The worked examples later illustrate the statistical calculations using explicitly hypothetical inputs.

**11.2 Dependence and multiple effects**

A single industrial study can report several assets, outcomes and time points. These estimates often share a comparator, programme or measurement system. Treating them as independent can overstate precision and give a prolific report more influence than an equally informative study with fewer reported outcomes.

The preferred handling depends on the question and available information. A review can select one pre-specified outcome per study, combine compatible within-study results using their covariance, fit a multilevel

model or use a suitable robust-variance approach. Each option has data and sample-size requirements. None should be applied as a cosmetic adjustment after pooling incompatible outcomes.

Shared control groups create another dependence structure. If several intervention arms are compared with the same baseline, the baseline information should not be counted repeatedly as independent evidence. The extraction record should preserve the comparison network and the number of independent units.

In industrial cases with one plant and many sensor observations, the deployment is still one operating context. Technical test-set size may support a model-performance estimate, while the operational programme effect remains based on one implementation. The review should report both counts and avoid using the larger technical count as the sample size for a financial effect.

### 11.3 Moderator analysis and transferability

Moderator analysis asks whether effect estimates vary with recorded study characteristics. Candidate moderators in predictive maintenance include comparator strategy, asset class, operating exposure, intervention scope, response resources and observation horizon. Their selection should be grounded in the proposed mechanism and pre-specified where possible.

Study-level associations can be confounded. A larger effect in one industry may coincide with weaker designs, different baselines or different cost boundaries. Meta-regression does not automatically separate these explanations, especially with few studies. The analysis should avoid fitting a long list of moderators to a small evidence set.

Transferability requires more than a moderator coefficient. A receiving site needs the conditions that make the intervention actionable: relevant signals, compatible assets, response time, spare parts, decision authority and an economic consequence worth changing. The review's transferability ledger records those conditions directly.

For GCC applications, the paper proposes a receiving-site assessment rather than a regional multiplier. The assessment compares the source intervention with the target operating system and identifies which assumptions require local evidence. A published saving can inform a hypothesis or scenario while remaining an estimate from another setting.

## 12. Risk of bias and reporting quality

Risk of bias concerns whether the study's estimate systematically differs from the effect it intends to estimate. Reporting quality concerns whether the information needed to assess that estimate is available. A clearly reported study can have substantial bias, while a poorly reported study can leave bias unresolved. The review should keep these judgements distinct.

The proposed industrial assessment draws conceptually on outcome-specific, design-aware approaches such as ROBINS-I [15]. That tool was developed for non-randomised intervention studies in health research. The paper does not claim a validated industrial adaptation. It uses the underlying questions about confounding, selection, measurement and reporting to structure a transparent assessment.

For before-and-after maintenance cases, key questions include whether a deterioration episode triggered adoption, whether other changes occurred, whether exposure was comparable and whether outcomes were measured consistently. For surveys, selection and strategy classification are central. For technical evaluations, leakage, label quality and independence of the test set are important.

The assessment should be supported by source locations and reasons. A numerical quality score can conceal a serious problem by averaging it with strengths elsewhere. The paper therefore proposes domain-level judgements and an explicit explanation of how each limitation affects the synthesis.

### 12.1 Regression to the mean and concurrent changes

A programme may be introduced after an unusually poor period. Even without an effective intervention, later performance can move closer to the plant's usual level. A simple before-and-after comparison can therefore overstate improvement if it uses an exceptional baseline. The review should examine why and when adoption occurred.

Concurrent changes can also affect the outcome. An overhaul, new operator training, better raw materials, a different product schedule or a capacity upgrade may accompany the programme. The study may evaluate the combined package legitimately, but attribution should match that package. An isolated AI claim requires additional evidence.

A longer pre-period can help describe baseline variability, while an appropriate comparison group can support a stronger design. Neither is automatically sufficient. The relevant assumptions concern what would have happened without the programme and whether the comparison captures that trajectory.

The extraction instrument therefore includes an adoption-trigger field and a concurrent-intervention register. These fields are often more informative for interpretation than the algorithm's name. They connect the reported outcome to the actual implementation history and help identify which studies can contribute to a common causal question.

**12.2 Selective publication and selective outcomes**

Successful deployments may be more likely to be written up than unsuccessful or inconclusive ones. Within a report, favourable outcomes may receive more attention than costs, false alerts or abandoned components. The review should consider both publication selection and selective outcome reporting.

Commercial case studies can provide valuable operational detail, but their purpose and selection process differ from a prospectively designed evaluation. The review should record sponsorship, author affiliation, customer identity where disclosed and the availability of an evaluation protocol. These characteristics inform interpretation without establishing bias by themselves.

Funnel plots and statistical tests have limited meaning with small, heterogeneous evidence sets. A symmetrical plot would not establish absence of publication bias, and an asymmetrical plot can arise for several reasons. Cochrane's guidance on missing evidence provides a methodological reference for these issues [16].

The proposed systematic-review protocol includes searches for discontinued, null and adverse outcomes, as well as positive case reports. It also records whether implementation cost, unnecessary maintenance and operational disruption were measured. A balanced economic synthesis requires the full consequence profile, not only the headline benefit.

**13. A protocol for a comprehensive systematic review**

The bounded audit can be extended through a prospectively specified search and selection process. The protocol should define the population, intervention, comparator, outcome families, study designs, publication types, languages and time period before screening. PRISMA 2020 and PRISMA-S support transparent reporting of that process [6,7].

The search should combine terms for predictive or condition-based maintenance, industrial AI and relevant operational or economic outcomes. It should include bibliographic databases appropriate to engineering and management, citation searching and a defined grey-literature strategy. The exact platforms, search strings, dates and limits should be retained. Search counts are reported only after the searches are actually executed.

Screening should distinguish technical benchmark papers from studies reporting deployment outcomes. Both can be retained in separate evidence maps if relevant to the review question. A technical-only study should not be misclassified as an operational intervention evaluation because its introduction mentions downtime savings.

The protocol should also define study-family linkage and full-text exclusion reasons. Records, reports and studies are counted separately. The final flow diagram should reconcile those counts. This avoids a common ambiguity in which several reports of one programme are presented as several independent deployments.

**13.1 Eligibility and screening decisions**

Eligibility should be applied at the level needed for the question. A paper may be eligible for the evidence map but contribute no poolable effect for a particular outcome. Excluding the entire paper because a standard error is missing can hide useful evidence about reporting and implementation.

The proposed primary operational question concerns deployed interventions in industrial settings with a described comparator or temporal comparison and an operational or economic outcome. Simulation-only studies and laboratory benchmarks can be mapped separately. Their inclusion in an operational-effect pool would require a clearly different estimand.

Screening decisions should retain the reason for exclusion and the information on which it was based. Ambiguous records can be resolved through full text or an author query where feasible. An unavailable full text should not be assumed to contain no relevant outcome. Its status remains unavailable within the review's access conditions.

For a completed systematic review, the number and role of screeners, disagreement process and any automation should be described accurately. The present paper proposes those procedures; it does not claim that

independent duplicate screening occurred for the bounded audit. Accurate method reporting is part of the evidence, not a cosmetic feature of the final manuscript.

**13.2 Extraction and reconciliation workflow**

The extraction form should be piloted on several different study designs. The purpose is to test whether the fields capture the information needed for synthesis without forcing ambiguous reports into false precision. Changes to the form should be recorded and applied consistently to earlier records.

Numerical extraction proceeds from source value to unit check, arithmetic check, boundary check and uncertainty check. The reviewer then determines the effect measure and pooling stratum. This order reduces the risk of choosing a convenient effect measure before understanding what the reported number represents.

Discrepancies should have a resolution record. A source correction, author clarification or explicit review decision can resolve some issues. Others remain unresolved and enter sensitivity analysis or narrative synthesis. The original value is preserved in all cases.

The final evidence table should allow a reader to trace a summary statement back to the source. A page or table locator, report identifier and calculation formula are usually sufficient for a compact paper, while the full extraction ledger can accompany it electronically. This combination supports readable reporting and reproducible detail.

**14. Hypothetical meta-analytic demonstrations**

The numerical demonstrations in this section are explicitly hypothetical. They illustrate statistical mechanics and do not represent additional maintenance studies. The examples use log ratios because the transformation makes proportional effects straightforward to combine and interpret under stated assumptions.

Suppose three independent studies of a genuinely comparable outcome report log ratios of −0.20, −0.40 and −0.60, with standard errors of 0.10, 0.20 and 0.30. The inverse-variance weights are 100, 25 and approximately 11.11. Their weighted mean is approximately −0.2694. Exponentiation gives a ratio of approximately 0.764, corresponding to a conditional reduction of about 23.6%.

The unweighted mean log ratio is −0.40, corresponding to a reduction of about 33.0%. The difference arises because the smallest hypothetical effect has the greatest precision. Neither calculation establishes which estimate is appropriate for a real evidence set; that depends on the estimand, variance validity and model assumptions.

The example shows why a simple average of headline percentages can be misleading. It ignores precision and can also differ from an average on the log scale. The appendices provide the inputs and calculations so that the distinction can be reproduced without relying on software defaults.

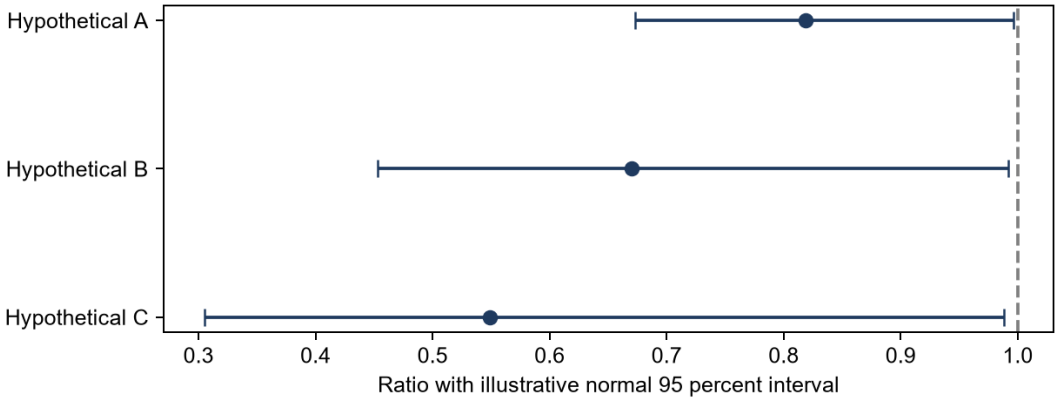



Figure 3. Explicitly hypothetical effect estimates and intervals. These are teaching examples and are not industrial study results.

**14.1 Heterogeneity sensitivity demonstration**

Continue the hypothetical example and assume, solely for demonstration, a between-study variance of 0.04. The random-effects weights become 1/(0.01+0.04), 1/(0.04+0.04) and 1/(0.09+0.04), or 20, 12.5 and approximately 7.69. The weighted mean becomes approximately −0.3388, corresponding to a ratio near 0.713 and a reduction near 28.7%.

The assumed heterogeneity reduces the dominance of the most precise study. This is a mathematical consequence of the weighting rule. It does not justify choosing a heterogeneity value that produces a preferred pooled result. In an empirical analysis, the variance estimator and uncertainty method must be stated and examined.

With few studies, uncertainty about between-study variation can be substantial. A pooled mean can appear stable while a prediction for a new setting remains wide. The review should distinguish the uncertainty around the mean from the variation expected across settings under the model.

The demonstration also reinforces the conceptual gate. Adding a heterogeneity term cannot make incompatible outcomes comparable. If one input represents downtime cost and another represents model accuracy, the weighted mean has no coherent interpretation regardless of how carefully the arithmetic is performed.

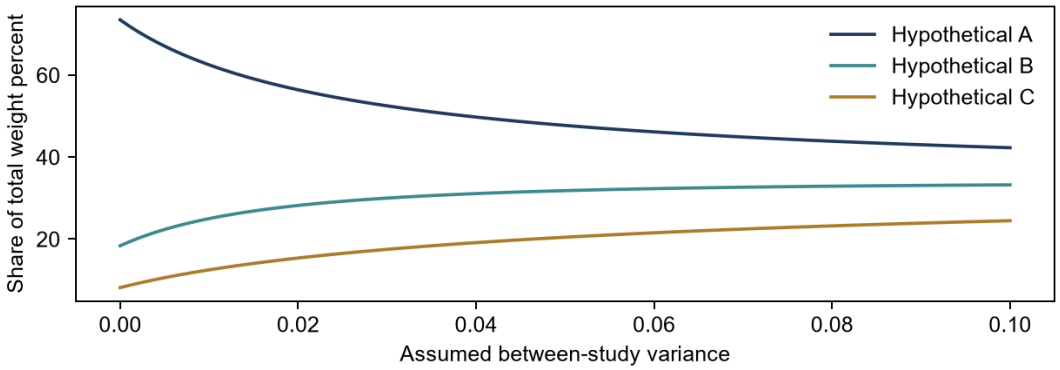



Figure 4. Weight sensitivity in the explicitly hypothetical example. Calculated from the formula in Appendix N.

**14.2 Hypothetical exposure adjustment**

Suppose an explicitly hypothetical plant records 20 failures over 10,000 operating hours before a programme and 12 failures over 8,000 hours after it. The raw event count falls by 40%. The event rates are 0.002 and 0.0015 per hour, giving a rate ratio of 0.75 and a rate reduction of 25%. The different answers arise from the change in exposure.

Under the simplified independent Poisson approximation, the standard error of the log rate ratio is the square root of 1/12 + 1/20, approximately 0.365. A conventional normal interval on the log scale would be broad. This illustrates that a substantial point reduction can coexist with considerable statistical uncertainty when event counts are small.

The example assumes comparable event definitions and independent counts. Recurrent failures, changing asset populations or overdispersion would require a different model or additional adjustment. The arithmetic is therefore an illustration of a stated model, not a recommended uncertainty estimate for every industrial case.

For evidence extraction, the lesson is to retain counts and exposure separately. If only the 40% headline reduction were stored, the reviewer could not reconstruct the exposure-adjusted comparison. The ledger's component fields preserve that option and make the resulting estimand explicit.

**14.3 Hypothetical dependence and sample size**

Suppose one plant reports outcomes for ten assets exposed to the same programme and operating environment. The ten effect estimates may share management, maintenance staff, utilities and a comparison period. Treating them as ten independent deployments can overstate the breadth and precision of the evidence.

A simplified cluster illustration uses the design effect $1+(m-1)\rho$, where m is the number of observations per cluster and $\rho$ is an assumed intracluster correlation. With m equal to 10 and $\rho$ equal to 0.20, the design effect is 2.8. Under the simplified assumptions, a nominal sample of 100 observations would have an effective size of about 35.7. This is an explicitly hypothetical demonstration, not an estimate for the examined studies.

The formula is a teaching device for one dependence structure. Real industrial data may have unequal cluster sizes, serial dependence and multiple levels. The appropriate model should follow that structure. A single effective-sample-size adjustment cannot resolve every dependence problem.

The review should therefore retain plant, programme and asset identifiers where possible. It should also distinguish technical sample size from operational replication. A million sensor readings can provide detailed

information about one deployment while still offering limited evidence about how the programme performs across independent industrial settings.

**15. Implications for industrial evaluation**

The evidence audit supports a practical reporting agenda. A deployment evaluation should define its comparison before implementation where feasible, retain a stable outcome dictionary and record the action pathway from alert to intervention. It should report operating exposure and adverse consequences alongside favourable outcomes.

The evaluation should also separate technical and economic claims. Technical validation can establish performance on a specified dataset. Operational evaluation examines whether the information changes maintenance decisions and outcomes. Economic evaluation values those consequences within a defined cost boundary. A complete programme can address all three, with the evidence for each stated separately.

For procurement, published cases are most useful when they reveal the conditions of success and the resources required. A headline percentage without asset scope, comparator and implementation cost provides limited support for a receiving site's business case. The estimand ledger can be used as a structured set of questions when assessing such evidence.

For research, better reporting would increase the possibility of meaningful synthesis. Comparable event definitions, exposure measures, uncertainty information and study-family identifiers would allow future reviews to move beyond disconnected percentages. The paper's proposed instruments are intended to support that improvement.

**15.1 Evaluation design for a receiving plant**

A receiving plant should begin with the decision it expects the programme to change. The relevant population may be a set of assets with a particular failure mechanism. The outcome should reflect the operating consequence of interest and the exposure over which it is measured. A broad claim about all plant reliability can be inappropriate for a narrow initial deployment.

The evaluation plan should identify the comparator, the observation horizon and the intervention bundle. It should record why assets were selected. Selecting the worst-performing assets can be commercially sensible while creating regression-to-the-mean concerns for evaluation. A credible design can account for that selection or limit its causal claims accordingly.

The plant should preserve alerts, reviews, actions and outcomes with linked identifiers. False alerts, rejected recommendations and delayed actions are relevant evidence. They can explain why strong technical performance does not translate into the expected operational result. They also identify opportunities to improve the operating process.

The financial appraisal should use the plant's own production constraints, recovery capacity and contribution margins. An international case can inform a scenario or an implementation question. It should not replace local measurement of the consequence that makes the programme economically valuable.

**15.2 Reporting a complete consequence profile**

A complete consequence profile includes programme cost, monitoring and review effort, planned interventions, unnecessary work, unplanned failures, production effects and financial outcomes. The profile should identify which values are observed and which are estimated. It should preserve the same horizon and boundary across the comparison.

Unnecessary interventions are particularly relevant to an alert-driven system. A reduction in unplanned failures can coexist with increased planned downtime and parts consumption. Whether the trade-off is favourable depends on cost and consequence. Reporting only failures can conceal that trade-off.

Programme continuity also matters. A successful pilot may use intensive support that is not sustained during routine operation. The evaluation should distinguish pilot conditions from the intended operating model. Expansion to more assets can change alert volume, response capacity and costs, so the effect should not be assumed to scale linearly.

The review's proposed reporting minimum therefore includes implementation resources and sustained-use evidence. These fields support interpretation of both positive and disappointing outcomes. They also make it easier for a future synthesis to group genuinely comparable interventions and explain variation between them.

## 16. Limitations and conclusion

The empirical component of this paper is a bounded audit of a purposively selected evidence set. It does not estimate the prevalence of reporting problems in the wider literature. The identified discrepancies are specific to the examined values and sources. A comprehensive systematic search could identify additional studies with stronger or different evidence.

The review also relies on published information. It does not have the underlying event-level or financial records needed to resolve every inconsistency. Recalculation can establish what follows from printed values, while leaving the accuracy of those values unresolved. The paper preserves that distinction.

The meta-analytic framework and hypothetical demonstrations establish methods, not an empirical industry effect. Future pooling would require compatible independent estimates, adequate uncertainty information and a defensible synthesis question. The proposed estimand ledger, extraction rules and bias assessment provide a structure for making that decision transparently.

The central conclusion is that reported predictive-maintenance outcomes become more useful when their measurement and comparison are preserved. The source audit shows how arithmetic, denominators and analytical boundaries can change interpretation. A rigorous synthesis should build from those foundations, reporting study-specific evidence where appropriate and pooling only a coherent set of effects.

## Appendix A. Evidence set and report-family register

The register defines the bounded evidence set examined in this paper. It separates a study unit from its reports and identifies the role each source plays. The register does not claim that these sources exhaust the literature. It provides a reproducible boundary for the numerical audit and the cross-study discussion.

| Study unit | Report source | Role in this paper |
|---|---|---|
| **US manufacturing survey** | Thomas and Weiss 2020 NIST AMS 100-34 | Survey design economic categories and response context |
| **US manufacturing survey** | Thomas and Weiss 2021 journal publication | Reported maintenance-strategy group comparison |
| **Tube-filling deployment** | Natanael and Sutanto 2022 | Before-and-after OEE and downtime audit |
| **RPA and ML deployment** | Patrício Varela and Silveira 2025 | Reliability and cost reporting audit |
| **Hot-rolling run-out table** | Yun Choi and Lee 2025 | Technical operational and economic boundary audit |

Four study units and five reports are represented. The two NIST publications are linked as one evidence family. The three industrial cases remain separate units. The review does not convert individual outcomes or component categories into additional independent deployments.

Source identity is preserved through title, author, year, DOI or institutional identifier, and the examined file. A broader review should add publication version, corrections, retractions where applicable, funding and competing-interest disclosures. Those fields should be extracted from the source rather than inferred from author affiliation.

The role column also limits interpretation. The NIST report contributes survey context, while a case table contributes a particular numerical comparison. A source can be useful for one purpose without supporting another. The synthesis should identify that purpose instead of assigning a single broad credibility label to the entire publication.

## Appendix B. Outcome-level extraction ledger

Each row in the full extraction ledger should represent one outcome comparison. The fields below preserve the information needed to decide whether the row belongs in a descriptive map, a structured synthesis or a statistical pool. The ledger is designed to retain uncertainty rather than force every source into a complete numeric record.

| Field | Required entry | Reason |
|---|---|---|
| **Study family** | Stable project or survey identifier | Avoid duplicate evidence |

| Field | Required entry | Reason |
|---|---|---|
| **Report identifier** | Publication and version | Trace the source |
| **Locator** | Page table figure or section | Verify the extracted value |
| **Population** | Assets plants or establishments | Define the unit represented |
| **Intervention** | Implemented programme components | Preserve the treatment bundle |
| **Comparator** | Actual prior or comparison practice | Define the contrast |
| **Outcome family** | Technical decision reliability production financial | Prevent incompatible pooling |
| **Outcome definition** | Numerator denominator and unit | Preserve meaning |
| **Observation window** | Dates and duration | Distinguish annualisation |
| **Exposure** | Operating hours units or other denominator | Adjust comparisons where justified |
| **Before and after values** | Original reported numbers | Reproduce arithmetic |
| **Reported change** | Source's own effect statement | Preserve the published claim |
| **Recalculated change** | Formula and result | Separate review calculation |
| **Uncertainty** | Variance interval or raw inputs | Determine weighting eligibility |
| **Evidence type** | Observed source-estimated or hypothetical | Keep result status explicit |
| **Discrepancy** | Nature and resolution status | Support sensitivity analysis |

The ledger should store missing fields as missing. A publication year does not substitute for an observation period, and a technical test-set size does not substitute for the number of independent plants. These distinctions become critical when weights or confidence intervals are calculated.

Where a source presents several alternative values, each can be stored with a shared outcome identifier and a version or scenario code. The review then records which value is used for a particular analysis and why. This approach preserves the original evidence while making the analytical choice transparent.

**Appendix C. Tube-filling downtime reconciliation**

The following values are transcribed from Table 11 of Natanael and Sutanto [3]. The totals in the final two rows distinguish the sum of the displayed categories from the printed total. The source's underlying event records are not available in this review, so the discrepancy is retained rather than resolved by assumption.

| Downtime category | Before minutes | After minutes |
|---|---|---|
| **Leaked tube sealing** | 474 | 544 |
| **Perforated tube sealing** | 152 | 13 |
| **Wrinkled tube sealing** | 228 | 104 |
| **Unstable tube weight** | 327 | 62 |
| **Filling pump no movement** | 425 | 65 |
| **Filling pump overload** | 229 | 7 |
| **Leak in seal rotary valve** | 166 | 0 |
| **Coding station overload** | 65 | 0 |
| **Runny nozzle** | 329 | 106 |

| Downtime category | Before minutes | After minutes |
|---|---|---|
| **Sum of displayed categories** | 2395 | 901 |
| **Printed total in source** | 2520 | 901 |

Using the row sum, the reduction is (2395−901)/2395, or 62.38%. Using the printed total, it is (2520−901)/2520, or 64.25%. The difference between the two baseline totals is 125 minutes. Both percentages appearing in the source can therefore be reproduced from displayed information, but they use different baselines.

This reconciliation has a specific interpretation. It establishes that the table and narrative are not numerically consistent on the baseline total. It does not establish whether a category was omitted, a value was mistyped or another record explains the difference. A reviewer should retain both alternatives and seek clarification before using one as a precise pooled estimate.

The category with increased downtime also remains visible. An aggregate reduction does not imply improvement in every failure mode. That observation can inform a mechanism discussion without treating each category as an independent study.

**Appendix D. OEE and percentage-point audit**

Table 10 of the tube-filling study reports the values below [3]. The absolute change is calculated by subtracting the before percentage from the after percentage. The relative change divides that difference by the before value. Calculations use the displayed values, so rounding in the original data can affect the final decimal.

| Measure | Before | After | Absolute change |
|---|---|---|---|
| **Availability rate** | 75.10% | 86.30% | 11.20 percentage points |
| **Performance rate** | 97.40% | 99.60% | 2.20 percentage points |
| **Quality rate** | 99.00% | 99.50% | 0.50 percentage points |
| **Overall equipment effectiveness** | 72.42% | 85.53% | 13.11 percentage points |

The OEE relative increase is approximately 18.10%. The availability relative increase is approximately 14.91%. These values differ from their percentage-point changes because the baseline is below 100%. The review reports the type of change explicitly rather than use an unqualified percentage-improvement label.

OEE is a composite. A reviewer should preserve the source's calculation and aggregation convention. The product of average component rates need not equal the average of period-level OEE values when the averaging scheme differs. A small discrepancy between those quantities should therefore be investigated before being labelled an arithmetic error.

For a future synthesis, absolute OEE changes can be grouped only where the definitions and operating exposure are sufficiently compatible. Relative changes can be displayed separately. The choice should be made in the protocol and should retain the source's baseline so that readers can understand the scale of the change.

This appendix illustrates a general reporting rule: retain original levels whenever possible. A percentage change alone discards information about baseline performance, ceiling effects and the practical size of the improvement.

**Appendix E. RPA case arithmetic and financial boundaries**

The selected values below come from Tables 14 and 15 of Patrício and colleagues [4]. Recalculations are based on the printed before-and-after values. The source describes company-provided operating and cost information. The review does not treat the table as an independent audit of the company's underlying records.

| Measure | Before | After | Recalculated change |
|---|---|---|---|
| **MTBF hours per event** | 23.47 | 70.4 | About 199.96% increase |
| **MTTR hours** | 12 | 4 | 66.67% reduction |
| **Unplanned events per month** | 15 | 5 | 66.67% reduction |

| Measure | Before | After | Recalculated change |
|---|---|---|---|
| **Maintenance cost EUR per month** | 4000 | 2500 | 37.50% reduction |
| **Downtime cost EUR per month** | 3500 | 1000 | 71.43% reduction |
| **Operating cost EUR per month** | 30000 | 25000 | 16.67% reduction |

The displayed MTBF pair implies an approximate tripling, while the source states a 100% improvement. The audit records this as a discrepancy between the printed values and their stated relative change. The rounded inputs explain the small difference from exactly 200%; they do not explain the difference from 100%.

The cost rows have different scopes. They should not be summed as percentages, and their monetary differences should not be added without establishing whether the categories overlap. A total operating-cost measure may already include maintenance and downtime-related expenditure. The ledger therefore stores each row's boundary separately.

For synthesis, the MTBF, event-rate and cost outcomes belong to different strata. They share one deployment and comparison history, so they are dependent evidence. Their numerical consistency can be assessed separately from the design's ability to identify a causal programme effect.

**Appendix F. Hot-rolling analytical bases**

The hot-rolling study presents several related calculations [5]. The table below preserves their locations and roles. It does not assume that the differing values refer to the same population or scenario. That relationship would need to be clarified before constructing a single outcome estimate.

| Source location | Reported input or outcome | Interpretation retained |
|---|---|---|
| **Table 16** | 2.3 downtime hours per year | Historical basis for profitability analysis |
| **Table 16** | USD 551000 annual TBM cost | Cost basis under stated replacement-cycle assumption |
| **Table 17** | USD 26000 opportunity cost | Source economic calculation using 2.3 hours |
| **Table 17** | USD 500000 comparison maintenance cost | Cost scenario used in one benefit calculation |
| **Table 18** | 11.8 versus 2.3 downtime hours per year | Reported PM versus PdM comparison |
| **Table 18** | USD 551000 versus USD 474000 maintenance cost | Reported comparative cost pair |

The Table 18 downtime pair implies an 80.51% reduction using the displayed values. Its cost pair implies approximately 13.97%. The Table 17 cost calculation uses a different comparison value. The review keeps these results separate because their analytical bases differ in the published presentation.

The consequence for synthesis is practical. A reviewer cannot select the largest saving and label it the study's unique effect without explaining the choice. A source can contain a historical loss estimate, a projected benefit and a comparative operating result. Each belongs to its own ledger row with its own status.

The appropriate next evidence would be a reconciliation of observation periods, plant scope and scenario assumptions. Until that information is available, the review can discuss the study's technical and analytical contribution while limiting the precision of any cross-study operational comparison.

**Appendix G. Outcome taxonomy**

The taxonomy below groups outcomes by the stage of the intervention they describe. It is a proposed organising device for industrial evidence synthesis. It should be applied to the actual definition in the source, because familiar labels can be used inconsistently.

| Family | Example measures | Required boundary |
|---|---|---|
| **Technical discrimination** | Precision recall F1 accuracy | Labels test set threshold and independent unit |
| **Prognostic timing** | Warning lead time remaining-life error | Event horizon and censoring |

| Family | Example measures | Required boundary |
|---|---|---|
| **Decision process** | Reviewed alerts accepted recommendations | Eligible decision episodes |
| **Execution** | Timely completed interventions | Action deadline and completion evidence |
| **Reliability frequency** | Failures per operating exposure | Event definition and exposure |
| **Reliability duration** | Repair time downtime hours | Asset line or plant boundary |
| **Production** | Saleable output yield OEE | Product schedule and quality definition |
| **Financial expenditure** | Repair maintenance operating cost | Included resources and price basis |
| **Financial consequence** | Lost contribution avoided loss | Counterfactual recovery and horizon |
| **Environmental** | Energy intensity emissions waste | Physical boundary and factors |
| **Sustained use** | Continued coverage adoption | Eligible asset population and period |
| **Adverse effects** | False interventions extra downtime | Complete consequence profile |

A study can contribute to several families. The synthesis should retain the relationships between those outcomes without collapsing them into a single score. A composite score would require a justified weighting scheme and could conceal an adverse trade-off.

The taxonomy also helps identify gaps. A literature may contain many technical evaluations and few operational comparisons. That pattern is a descriptive finding about the examined evidence, provided the search and selection support it. The bounded review here illustrates the distinction without estimating its prevalence across the entire literature.

For decision use, the relevant family depends on the question. Selecting a model may require technical metrics. Funding a deployment requires evidence on actionability and economic consequence. A good evidence synthesis makes those differences visible.

**Appendix H. Estimand compatibility matrix**

Compatibility is assessed before pooling. The matrix below identifies the dimensions that must be compared. A mismatch does not necessarily exclude a study from the review; it can place the study in a different stratum or require a narrower synthesis question.

| Dimension | Compatible basis sought | Example of a consequential mismatch |
|---|---|---|
| **Population** | Comparable assets or industrial units | Selected components versus whole plants |
| **Intervention** | Similar operating bundle | Algorithm test versus deployed response programme |
| **Comparator** | Clearly defined alternative practice | Reactive baseline versus mature preventive programme |
| **Outcome** | Same substantive construct | Downtime duration versus downtime cost |
| **Unit** | Consistent physical or financial scale | Event count versus exposure-adjusted rate |
| **Horizon** | Comparable observation and reporting periods | Short pilot versus annual projection |
| **Design** | Similar target effect | Cross-sectional association versus controlled intervention |
| **Uncertainty** | Defensible sampling variance | Observed standard error versus invented precision |
| **Independence** | Distinct evidence units | Multiple reports of one deployment |

| Dimension | Compatible basis sought | Example of a consequential mismatch |
|---|---|---|
| **Cost boundary** | Comparable included resources | Gross saving versus net programme benefit |

The matrix should be accompanied by a written decision for each proposed pool. The decision states what the pooled estimate would mean and why the included studies can answer that question. A statistical heterogeneity test is considered after this conceptual assessment.

Some differences can be harmonised. Hours and minutes can be converted exactly. Currency values can be converted under a stated price and exchange-rate convention. Other differences cannot be removed through arithmetic. A projected avoided loss and a realised expense reduction remain different evidence types after unit conversion.

When compatibility is incomplete, a structured display can still be valuable. It can show study-specific effects with their definitions and explain the reasons for separate grouping. This preserves useful evidence without giving a misleading precision to an overall average.

**Appendix I. Proposed systematic search specification**

This appendix describes a prospective search structure for a comprehensive review. It is not a record of searches already completed. A future review should retain exact database syntax, platform, date, limits and exported records for each executed search, following transparent search-reporting principles [7].

| Concept block | Candidate terms | Purpose |
|---|---|---|
| **Maintenance approach** | predictive maintenance condition-based maintenance prognostics | Identify relevant intervention families |
| **Industrial analytics** | industrial AI machine learning anomaly detection | Capture analytical implementations |
| **Deployment** | plant factory manufacturing industrial field implementation | Distinguish operational settings |
| **Reliability outcomes** | failure downtime availability repair MTBF MTTR | Capture operating consequences |
| **Production outcomes** | throughput yield OEE production loss | Capture production effects |
| **Economic outcomes** | cost savings return investment profitability | Capture financial analysis |
| **Adverse outcomes** | false alarm unnecessary maintenance abandoned deployment | Reduce favourable-outcome selection |

The search should avoid requiring every outcome term in the title or abstract, because relevant operational results may appear only in the full text. A broad intervention-and-setting search can be supplemented by targeted outcome searches and citation tracing. The trade-off between sensitivity and screening workload should be documented.

Engineering and management databases may index different parts of the literature. Grey literature can contain deployment evidence absent from journals, while raising additional provenance and selection questions. The protocol should define which source types are eligible and how they are assessed.

The final search record should distinguish retrieved records from unique reports and study units. It should also identify updates and corrections. Search completeness remains a methodological claim that requires the executed record; it cannot be inferred from the length of a bibliography.

**Appendix J. Screening and exclusion framework**

The screening framework separates review eligibility from pooling eligibility. A study can be relevant to the evidence map while lacking a compatible effect estimate. This distinction avoids discarding useful information merely because it cannot enter a statistical model.

| Question | Inclusion implication | Record required |
|---|---|---|
| **Industrial setting present** | Eligible for deployment map | Setting and operating context |

| Question | Inclusion implication | Record required |
|---|---|---|
| **Intervention described** | Assess programme scope | Components and implementation stage |
| **Operational outcome reported** | Eligible for operational synthesis consideration | Outcome definition and locator |
| **Comparator identified** | Classify comparison design | Before period or comparison group |
| **Technical-only evaluation** | Separate technical stratum | Test design and metric |
| **Simulation-only outcome** | Separate modelling stratum | Model assumptions and validation basis |
| **Economic projection** | Separate appraisal stratum | Inputs horizon and cost boundary |
| **Duplicate report** | Link to study family | Preferred source by outcome |
| **Unclear identity** | Retain overlap flag | Evidence supporting uncertainty |
| **Missing full text** | Record access limitation | Attempt and available metadata |
| **Incompatible pooled outcome** | Retain in review where relevant | Specific reason for non-pooling |

Exclusion reasons should be specific and mutually understandable. An entry such as poor quality is too vague to reproduce. A reason such as no deployed operational outcome or duplicate report of the same programme identifies the actual decision. More than one limitation can be recorded, with a primary exclusion reason for flow accounting.

For a completed review, the screening procedure should state how many people assessed records and how disagreements were resolved. Automation can support retrieval or prioritisation, but its role should be described accurately. This paper supplies the proposed decision framework and does not claim that independent duplicate screening was performed for the bounded source audit.

**Appendix K. Bias and reporting assessment domains**

The proposed domains below organise an industrial appraisal of outcome validity. They draw on design-aware risk-of-bias reasoning while retaining the need for an industrially justified application [15]. A domain judgement should include a reason and source locator. The table is not a validated numerical scoring instrument.

| Domain | Principal question | Evidence sought |
|---|---|---|
| **Selection** | Why were these assets or firms included | Eligibility and adoption trigger |
| **Confounding** | What else changed with the programme | Concurrent intervention and operating history |
| **Intervention classification** | When and how was exposure established | Deployment and routine-use records |
| **Outcome measurement** | Was the definition stable across comparison | Metric dictionary and collection method |
| **Missing outcomes** | Which periods assets or events are absent | Completeness and attrition record |
| **Selective reporting** | Were outcomes chosen after results were known | Protocol and full consequence profile |
| **Technical leakage** | Could test information enter model development | Split strategy and temporal separation |
| **Dependence** | Are observations treated at the correct level | Plant asset event and sample identifiers |
| **Numerical consistency** | Do levels totals and percentages reconcile | Independent arithmetic |

| Domain | Principal question | Evidence sought |
|---|---|---|
| **Economic boundary** | Are costs and benefits non-overlapping | Component and counterfactual bridge |

Reporting completeness is recorded separately. Missing information can leave a domain unresolved rather than establish that the study necessarily has a particular flaw. The synthesis should state how unresolved domains affect confidence in the specific result.

The assessment should also be outcome-specific. A paper may report a well-defined technical test and a weakly supported financial projection. Assigning one global score would obscure that difference. The estimand ledger links the domain judgement to the outcome and comparison it concerns.

Finally, the review should avoid treating sponsorship or publication venue as a substitute for methodological assessment. These characteristics are recorded, but the judgement rests on the design, measurement and reporting relevant to the claimed effect.

**Appendix L. Effect-measure selection guide**

The guide links the outcome's structure to a candidate effect measure. It does not prescribe one universal statistic. The final choice must match the study design, available uncertainty information and synthesis question. Raw components should be retained so that transformations can be checked.

| Outcome structure | Candidate measure | Required information |
|---|---|---|
| **Same-scale continuous outcome** | Mean difference | Means variability and independent sample sizes |
| **Positive continuous outcome** | Log response ratio | Means and appropriate variance basis |
| **Event counts with exposure** | Log rate ratio | Counts exposure and dependence assumptions |
| **Binary event outcome** | Risk ratio or risk difference | Events and eligible units |
| **Paired continuous change** | Mean change or paired effect | Before after variability and covariance |
| **Percentage-valued measure** | Percentage-point difference | Consistent definition and scale |
| **Time-to-event outcome** | Hazard-based or survival measure | Censoring and follow-up information |
| **Cost outcome** | Defined cost difference or ratio | Boundary currency price basis and uncertainty |
| **Modelled economic benefit** | Scenario result | Model inputs and assumptions |
| **Technical classification** | Metric-specific synthesis | Labels threshold prevalence and test design |

The same source may support several measures. The protocol should identify the primary measure and retain alternatives for sensitivity analysis. Choosing the measure that produces the largest apparent benefit would create selective interpretation.

Transformations should also preserve direction. A ratio below one is favourable for failure rate but unfavourable for availability. A forest plot should state the direction clearly and avoid reversing some effects without documenting the transformation. The source's original units remain in the extraction ledger.

When the required information is missing, the outcome can remain in the evidence map. A transparent descriptive summary is preferable to an effect estimate whose variance or denominator has been invented. The absence of a poolable statistic is itself relevant to the reporting agenda for future deployment studies.

**Appendix M. Hypothetical inverse-variance calculation**

This appendix reproduces the common-effect demonstration in Section 14. The three study labels below are hypothetical teaching examples. They are not publications or observed maintenance deployments. The

outcome is assumed to be conceptually identical and the estimates independent solely to isolate the weighting calculation.

| Example | Log ratio | Standard error | Weight |
|---|---|---|---|
| **A** | -0.20 | 0.10 | 100.0000 |
| **B** | -0.40 | 0.20 | 25.0000 |
| **C** | -0.60 | 0.30 | 11.1111 |
| **Total** | Not applicable | Not applicable | 136.1111 |

The weighted numerator is −36.6667. Dividing by 136.1111 gives −0.2694. Exponentiation gives approximately 0.7638. Under the stated direction of benefit, the corresponding reduction is approximately 23.62%. The standard error of the weighted mean under this simple model is the square root of 1/136.1111, approximately 0.0857.

A conventional normal 95% interval on the log scale is the weighted mean plus or minus 1.96 times that standard error. Exponentiating the endpoints gives the interval on the ratio scale. The calculation is conditional on the assumed variances, independence and common-effect model. It should not be transferred to the bounded industrial evidence set in this paper.

The example makes the role of precision visible. Example A receives about 73.5% of the total weight, B about 18.4% and C about 8.2%. An unweighted average would give each one-third. The choice of weights therefore has a substantive effect on the result even before heterogeneity or bias is considered.

An empirical review should verify that the reported standard errors correspond to the same estimand as the effect values. A standard deviation, confidence-interval width or model-fit error cannot be inserted as a standard error without the appropriate derivation.

**Appendix N. Hypothetical heterogeneity sensitivity**

The table adds an assumed between-study variance to the same hypothetical inputs. It demonstrates how the random-effects weighting formula changes influence. The values of $\tau^2$ are selected for illustration and are not estimated from the industrial studies reviewed in this paper.

| Assumed between-study variance | Weight A | Weight B | Weight C |
|---|---|---|---|
| **0.00** | 100.0000 | 25.0000 | 11.1111 |
| **0.01** | 50.0000 | 20.0000 | 10.0000 |
| **0.04** | 20.0000 | 12.5000 | 7.6923 |
| **0.09** | 10.0000 | 7.6923 | 5.5556 |

The formula is $w_i = 1/(v_i+\tau^2)$. As the assumed between-study variance increases, the difference between the weights becomes smaller. This follows because the common added component becomes more important relative to the individual sampling variances.

At $\tau^2$ equal to 0.04, the weighted mean log ratio is approximately −0.3388. Exponentiation gives approximately 0.7127, or a conditional reduction of 28.73%. This result differs from the common-effect demonstration because influence has shifted toward the larger hypothetical effects.

The demonstration does not imply that a random-effects model is more favourable or more conservative in every application. The direction of the change depends on the relationship between effect size and precision. In a different hypothetical dataset, the shift could reduce the estimated benefit or change the interpretation in another way.

An empirical analysis should state the heterogeneity estimator, interval method and sensitivity choices. It should also examine whether the studies belong to a coherent effect distribution. Statistical weighting is the final part of that decision, after conceptual compatibility, independence and risk of bias have been assessed.

**Appendix O. Hypothetical event-rate calculation**

The example records 20 events in 10,000 hours before an intervention and 12 events in 8,000 hours after it. These values are hypothetical. The event definition and asset population are assumed unchanged, and the independent Poisson approximation is used only to demonstrate a simple variance calculation.

| Quantity | Before | After |
|---|---|---|

| Quantity | Before | After |
|---|---|---|
| **Events** | 20 | 12 |
| **Operating exposure hours** | 10000 | 8000 |
| **Events per 1000 hours** | 2.00 | 1.50 |
| **Raw count ratio** | Not applicable | 0.60 |
| **Exposure-adjusted rate ratio** | Not applicable | 0.75 |

The log rate ratio is ln(0.75), approximately −0.2877. Its approximate standard error is the square root of 1/12+1/20, approximately 0.3651. A normal interval on the log scale is therefore much wider than the point estimate alone suggests. The event counts provide limited precision under the assumed model.

The raw count reduction is 40%, while the rate reduction is 25%. This difference is entirely due to the changed exposure in the example. A review that stores only the headline count reduction loses the information needed to distinguish improved reliability from reduced operating time.

Real industrial event processes may violate the approximation. Recurrent failures can cluster within assets, repairs can change subsequent risk, and common causes can affect several assets together. The review should use the study's appropriate uncertainty estimate where available and record its assumptions. It should not apply the simple formula to rounded average events without a defensible count and exposure basis.

The example supports a reporting minimum: event counts, exposure, population and event definition should accompany any rate claim.

**Appendix P. Hypothetical pairing sensitivity**

For repeated measurements on the same independent units, uncertainty in the mean change depends on the pre-post correlation. This appendix uses hypothetical standard deviations of 10 in both periods and 25 independent units. The variance of the individual change is 200−200r, and the standard error of the mean change is the square root of that variance divided by 25.

| Assumed correlation r | Change variance | Standard error of mean change |
|---|---|---|
| **0.00** | 200 | 2.8284 |
| **0.25** | 150 | 2.4495 |
| **0.50** | 100 | 2.0000 |
| **0.75** | 50 | 1.4142 |
| **0.90** | 20 | 0.8944 |

The table shows why the correlation cannot be ignored when constructing a paired effect. It also shows why an assumed correlation should be visible. Selecting a high value produces a smaller standard error in this example. An empirical review should not choose that value merely to increase the apparent precision of a result.

The number 25 represents independent units under the hypothetical design. It does not represent 25 sensor readings from one machine unless those readings satisfy the required independence assumptions. In practice, repeated industrial data can have several levels of dependence, including time, asset and plant.

If the source reports a paired standard error directly, the reviewer should preserve it and verify the outcome scale. If only marginal standard deviations are available, the protocol can specify a transparent sensitivity analysis. If even those components are missing, the outcome may remain descriptive.

The purpose is to make the uncertainty model inspectable. A point estimate calculated correctly from before-and-after means can still receive an inappropriate weight if its pairing and unit of analysis are mishandled.

**Appendix Q. Hypothetical clustering demonstration**

This illustration uses the simplified design effect 1+(m−1)ρ for equal-sized clusters. It assumes ten observations per cluster and shows how an assumed intracluster correlation changes the information

represented by a nominal total of 100 observations. The example is explanatory and is not an estimate for any reviewed study.

| Assumed intracluster correlation | Design effect | Approximate effective sample size |
|---|---|---|
| **0.00** | 1.00 | 100.00 |
| **0.05** | 1.45 | 68.97 |
| **0.10** | 1.90 | 52.63 |
| **0.20** | 2.80 | 35.71 |
| **0.30** | 3.70 | 27.03 |
| **0.50** | 5.50 | 18.18 |

The calculation divides the nominal sample size by the design effect. It illustrates why correlated observations provide less independent information than the same number of independent observations. It does not replace a model suited to unequal clusters, serial correlation or multiple hierarchical levels.

For predictive maintenance, the relevant cluster can be a plant, asset, production episode or programme. Several levels may coexist. The extraction should therefore retain the level at which the intervention is assigned and the level at which the outcome is measured. A large number of measurements at a lower level does not necessarily provide many independent intervention comparisons.

The practical implication for meta-analysis is that study precision should reflect the original design. If a paper reports an analysis that ignores clustering, the reviewer should consider whether an appropriate adjustment can be made from available information. Any assumed correlation belongs in a sensitivity analysis and should remain labelled as assumed.

This is also a reason to report the number of independent deployments separately from the total number of assets and sensor records.

**Appendix R. Structured synthesis without pooling**

A structured synthesis should define how studies are grouped, how outcomes are presented and how conclusions are drawn. SWiM provides reporting guidance for quantitative intervention evidence when a conventional effect-estimate meta-analysis is not used [8]. The industrial application here uses explicit design and outcome strata.

| Stratum | Included evidence type | Appropriate summary |
|---|---|---|
| **Technical evaluation** | Model metrics on specified test data | Metric and validation-design table |
| **Survey association** | Between-establishment comparisons | Group definition and association description |
| **Before-and-after deployment** | Operational change over periods | Source-linked effect with confounding limits |
| **Controlled deployment** | Defined comparison units and timing | Design-specific effect and uncertainty |
| **Economic model** | Conditional benefit calculation | Assumptions sensitivity and boundary |
| **Implementation experience** | Qualitative operational findings | Mechanism and context synthesis |

The synthesis should avoid vote counting based on statistical significance. A non-significant estimate can be imprecise rather than evidence of no effect, and a significant result can still be biased or operationally small. Direction, magnitude, uncertainty and design should be considered together.

The paper can present an effect-direction display when the rules are explicit, but it should not imply that the display estimates a common average. The number of favourable reports is also affected by selection and publication practices. A bounded set cannot establish the success rate of all deployments.

For the present evidence, the principal synthesis is a set of study-specific outcomes and a numerical-consistency audit. The conclusion concerns the interpretation and comparability of those reports. It does not convert them into a pooled industry-saving percentage. This form of synthesis preserves useful information while matching the strength and structure of the available evidence.

**Appendix S. Receiving-site transferability assessment**

The assessment compares a published intervention with a proposed receiving site. It is a research and appraisal instrument, not a scoring system that predicts savings. Each difference should be linked to the mechanism through which it could affect performance.

| Domain | Comparison question | Evidence needed locally |
|---|---|---|
| **Asset and failure mode** | Is the same deterioration mechanism relevant | Asset history and engineering assessment |
| **Signal availability** | Are required measurements reliable and timely | Data-quality and sensor assessment |
| **Operating regime** | Are loading product and environment comparable | Operating history |
| **Comparator practice** | What maintenance already occurs | Current programme and execution records |
| **Warning horizon** | Is there time for a useful response | Failure progression and response timing |
| **Action capability** | Are people parts and access available | Work execution and resource records |
| **Decision authority** | Who can approve intervention | Escalation and approval process |
| **Economic consequence** | What output or cost is actually at risk | Production recovery and margin bridge |
| **Scale** | Can review and response capacity support more assets | Workload and staffing model |
| **Sustained use** | Will the intended support model persist | Operating responsibilities and cost plan |

A difference can require adaptation rather than rejection. For example, a longer procurement lead time can make the same warning less actionable. A lower contribution margin can reduce the economic value of the same downtime change. The assessment should state the proposed adaptation and the evidence needed to evaluate it.

The published result can inform a scenario, but the scenario should use the receiving site's own exposure and cost boundary. A multiplier copied from a case study does not establish the local effect. The assessment identifies where local measurement would most improve the decision.

For GCC process industries, this approach preserves the relevance of international evidence while keeping the regional claim proportionate to the actual comparison.

**Appendix T. Minimum reporting set for future deployments**

The proposed minimum set aims to make future industrial evaluations easier to interpret and synthesise. It combines programme description, outcome measurement and economic boundary. The list is a proposed contribution of this paper and should be tested through use in real evaluations.

| Reporting item | Minimum information | Why it matters |
|---|---|---|
| **Population** | Assets plants and eligibility | Defines the evidence boundary |
| **Selection** | Why the units were chosen | Reveals adoption triggers and selection |
| **Intervention** | Technology workflow and execution resources | Describes the implemented bundle |
| **Comparator** | Actual alternative maintenance practice | Defines the effect contrast |
| **Timing** | Installation routine use and observation dates | Separates exposure from publication |
| **Outcome dictionary** | Numerator denominator unit and direction | Supports comparability |
| **Raw components** | Counts exposure before and after values | Permits recalculation |
| **Uncertainty** | Appropriate intervals or derivation inputs | Supports responsible weighting |
| **Concurrent changes** | Overhauls training schedules and assets | Informs attribution |
| **Costs** | Implementation recurring and response costs | Defines net economic consequence |
| **Adverse effects** | False interventions extra work and disruptions | Completes the consequence profile |
| **Sustained operation** | Coverage and performance beyond the pilot | Supports transfer and scale-up questions |
| **Source access** | Data availability and record locations | Supports verification |
| **Corrections** | Version and numerical reconciliation | Preserves the published record |

The set does not require disclosure of confidential plant data in full. Aggregated values, a clear dictionary and an appropriate access process can still support interpretation. The key is to preserve enough information to understand the result and its uncertainty.

The reporting burden should be considered during evaluation design. If the programme already links alerts, work orders and production events, many fields can be generated from operational records. Retrofitting the evidence after a successful case has been selected is more difficult and can increase selective reporting.

**Appendix U. Reproducibility and numerical audit protocol**

The numerical audit should be reproducible from retained source versions and explicit formulas. A source value, a reviewer calculation and an interpretation are stored separately. This structure permits a later correction without losing the original record.

| Stage | Retained object | Audit question |
|---|---|---|
| **Source capture** | Original report and identifier | Is the version identifiable |
| **Location** | Page table or section | Can the value be found |
| **Transcription** | Original value unit and label | Does it match the source |

| Stage | Retained object | Audit question |
|---|---|---|
| **Arithmetic** | Formula and calculated result | Does the stated change reconcile |
| **Aggregation** | Category totals and study-family map | Is evidence counted once |
| **Classification** | Outcome and evidence-type codes | Does the synthesis preserve meaning |
| **Uncertainty** | Variance inputs and assumptions | Is precision justified |
| **Pooling decision** | Compatibility and exclusion reasons | Is the combined estimand coherent |
| **Presentation** | Tables figures and narrative | Do claims match the ledger |
| **Revision** | Change log and prior values | Can altered conclusions be explained |

The audit should distinguish numerical verification from factual verification. Reproducing a percentage from printed values establishes the arithmetic. It does not establish that the printed values accurately reflect the underlying industrial records. Likewise, reproducing a meta-analysis from extracted effects establishes computation within the inputs, not the validity of the study designs.

For this paper, the principal completed checks are the source-linked recalculations shown in Appendices C through F and the explicitly hypothetical calculations in Appendices M through Q. The broader search, screening and pooling procedures are proposed for a future comprehensive review. Keeping those statuses explicit allows the paper to contribute a usable method without overstating the empirical work performed.

The resulting research package should preserve the extraction ledger, calculation code, source references and final tables. Subsequent updates can then add evidence through the same structure and show exactly which conclusions change.

**References**

[1] Thomas D and Weiss B. Economics of Manufacturing Machinery Maintenance A Survey and Analysis of U.S. Costs and Benefits. NIST AMS 100-34. 2020. https://doi.org/10.6028/NIST.AMS.100-34

[2] Thomas D and Weiss B. Maintenance Costs and Advanced Maintenance Techniques in Manufacturing Machinery Survey and Analysis. International Journal of Prognostics and Health Management. 2021;12(1). https://doi.org/10.36001/ijphm.2021.v12i1.2883

[3] Natanael D and Sutanto H. Machine Learning Application Using Cost-Effective Components for Predictive Maintenance in Industry A Tube Filling Machine Case Study. Journal of Manufacturing and Materials Processing. 2022;6(5):108. https://doi.org/10.3390/jmmp6050108

[4] Patrício L Varela L and Silveira Z. Proposal for a Sustainable Model for Integrating Robotic Process Automation and Machine Learning in Failure Prediction and Operational Efficiency in Predictive Maintenance. Applied Sciences. 2025;15(2):854. https://doi.org/10.3390/app15020854

[5] Yun JW Choi SW and Lee EB. Study on Energy Efficiency and Maintenance Optimization of Run-Out Table in Hot Rolling Mills Using Long Short-Term Memory-Autoencoders. Energies. 2025;18(9):2295. https://doi.org/10.3390/en18092295

[6] Page MJ McKenzie JE Bossuyt PM et al. The PRISMA 2020 statement an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71

[7] Rethlefsen ML Kirtley S Waffenschmidt S et al. PRISMA-S an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic Reviews. 2021;10:39. https://doi.org/10.1186/s13643-020-01542-z

[8] Campbell M McKenzie JE Sowden A et al. Synthesis without meta-analysis SWiM in systematic reviews reporting guideline. BMJ. 2020;368:l6890. https://doi.org/10.1136/bmj.l6890

[9] Cochrane Handbook for Systematic Reviews of Interventions. Chapter 6 Choosing effect measures and computing estimates of effect. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-06

[10] Cochrane Handbook for Systematic Reviews of Interventions. Chapter 10 Analysing data and undertaking meta-analyses. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-10

[11] Cochrane Handbook for Systematic Reviews of Interventions. Chapter 12 Synthesizing and presenting findings using other methods. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-12

[12] DerSimonian R and Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177–188. https://doi.org/10.1016/0197-2456(86)90046-2

[13] Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software. 2010;36(3):1–48. https://doi.org/10.18637/jss.v036.i03

[14] Higgins JPT Thompson SG Deeks JJ and Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–560. https://doi.org/10.1136/bmj.327.7414.557

[15] Sterne JAC Hernán MA Reeves BC et al. ROBINS-I a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919

[16] Cochrane Handbook for Systematic Reviews of Interventions. Chapter 13 Assessing risk of bias due to missing evidence in a meta-analysis. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-13

[17] Zonta T da Costa CA da Rosa Righi R de Lima MJ da Trindade ES and Li GP. Predictive maintenance in the Industry 4.0 A systematic literature review. Computers and Industrial Engineering. 2020;150:106889. https://doi.org/10.1016/j.cie.2020.106889

[18] Hoffmann MA and Lasch R. Unlocking the Potential of Predictive Maintenance for Intelligent Manufacturing a Case Study On Potentials Barriers and Critical Success Factors. Schmalenbach Journal of Business Research. 2025;77:27–55. https://doi.org/10.1007/s41471-024-00204-3

The four-study evidence set is defined in Appendix A. Methodological references support the synthesis framework; they are not additional deployment-effect observations. The numerical demonstrations are explicitly hypothetical and are excluded from the empirical evidence count.