JURNAL LOCUS: Penelitian & Pengabdian
Volume 5 No. 10 October 2026, pp 9415-9430
E-ISSN 2829-7334
DOI: 10.58344/locus.v5i10.6458

Effectiveness of Kaizen-Based Programmes in Improving Patient Safety Incident Reporting: A Systematic Review Protocol with Monte Carlo Validation

Andan Suryo Wardani*, Dyah Wiji Puspita Sari, Dwi Heppy Rochmawati, Muh. Abdurrouf
Universitas Islam Sultan Agung Semarang, Indonesia
Email: andansw89@gmail.com*, dyah.wiji@unissula.ac.id, dwiheppy@unissula.ac.id, abdurrouf@unissula.ac.id

Keywords:
Kaizen; Incident Reporting; Patient Safety; Reporting Propensity; Monte Carlo Simulation.

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Riviera Publishing

ABSTRACT

Incident reporting is a fundamental component of organizational learning in patient safety. However, Kaizen-based continuous improvement programs designed to strengthen incident reporting are often evaluated using report counts, which cannot distinguish improvements in reporting culture from increases in actual patient harm. This study aimed to develop and validate an analytical framework that differentiates these effects and establish a systematic review protocol for its application. The observed reporting rate was decomposed into three components: exposure, the true incident rate, and reporting propensity. A severity-stratified Reporting Propensity Index (RPI) was developed using partial identification bounds, alongside a behavioral reporting model, an assessment of bias in before-and-after study designs, and a sustainability half-life measure. Four Monte Carlo simulation studies encompassing 52 scenarios were conducted to evaluate estimator performance, and a systematic review protocol was developed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. Under the specified assumptions, the RPI estimator was unbiased and achieved nominal confidence interval coverage, whereas the conventional total-count estimator exhibited bias of up to −0.35 log units. When assumptions were violated, the partial identification bounds contained the true parameter in at least 97.4% of simulation replications. Uncontrolled before-and-after designs overestimated intervention effects, consistent with theoretical predictions, while reusing baseline-period data in moderator analyses increased the false-positive rate to 20.7%. These findings suggest that severity-stratified incident reporting, analyzed using the RPI framework, may provide hospitals with a more reliable measure of reporting propensity and support evidence-based decisions regarding the implementation and evaluation of Kaizen-based patient safety improvement programs.

INTRODUCTION

Preventable patient harm remains a major challenge in hospital care. A meta-analysis of 70 studies estimated that approximately one in 20 patients experienced preventable harm, with approximately 12% of these incidents resulting in permanent disability or death (Panagioti et al., 2019). A retrospective medical record review conducted across 11 hospitals in the United States identified adverse events in nearly one-quarter of hospital admissions,

approximately one-quarter of which were considered preventable (Bates et al., 2023). Learning from these events depends on effective patient safety incident reporting systems, which transform frontline observations into organizational knowledge and support continuous improvements in patient safety.

Despite their importance, incident reporting systems frequently demonstrate limited effectiveness in clinical practice. More than two decades after the publication of To Err Is Human, patient safety reporting has continued to face challenges related to fragmentation, underutilization, and insufficient integration with corrective actions (Mitchell et al., 2016). (Macrae, 2016) highlighted that reporting systems often prioritize data collection over meaningful analysis and provide inadequate feedback to healthcare professionals. Systematic reviews have consistently identified fear of blame, time constraints, ambiguous reporting criteria, complicated reporting procedures, and insufficient feedback as major barriers to effective incident reporting (Archer et al., 2017; Hamed & Konstantinidis, 2022; Oweidat et al., 2023). In Indonesia, hospital incident reporting systems have demonstrated limited alignment with the World Health Organization's recommendations for effective reporting systems, while healthcare professionals have reported persistent organizational, procedural, and cultural barriers (Dhamanti et al., 2020; Dhamanti et al., 2019). These challenges are further exacerbated by inadequate patient safety culture and limited psychological safety, which may discourage healthcare professionals from reporting incidents and near misses (Reis et al., 2018; ee et al., 2019); O'Donovan & McAuliffe, 2020; Grailey et al., 2021.

Kaizen, a Japanese management philosophy emphasizing continuous, incremental improvement and a foundational principle of Lean management, offers a potential approach to addressing these barriers. Daily Kaizen practices encourage employees to identify operational problems and propose improvements, while visual management boards and brief team meetings facilitate timely feedback and corrective action. The 5S workplace organization methodology supports operational efficiency, whereas the Plan–Do–Check–Act (PDCA) cycle provides a structured framework for implementing and evaluating improvements (Mazzocato et al., 2016; Kanamori et al., 2016; Rotter et al., 2019). Previous systematic reviews have demonstrated that Lean and Kaizen-based interventions can improve patient flow, reduce waiting times, and enhance selected healthcare process outcomes (Moraros et al., 2016; Tlapa et al., 2020). However, other reviews have reported inconsistent effectiveness, increased staff workloads, and challenges in sustaining improvements over time (Rees & Gauld, 2017; Hallam & Contreras, 2018; Leite et al., 2020; Mahmoud et al., 2021).

Three major research gaps emerge from the existing literature. First, systematic reviews have rarely examined patient safety incident reporting as the primary outcome of Kaizen-based interventions, as previous reviews have generally considered patient safety alongside broader organizational and clinical outcomes (Akmal et al., 2020; Marsilio & Pisarra, 2021). Second, evaluations of incident reporting interventions face a fundamental measurement challenge, referred to as the reporting paradox, whereby an increase in reported incidents may indicate either improved reporting behavior or an actual increase in patient harm. Conversely, a reduction in reported incidents may reflect improved patient safety or decreased willingness to report adverse events. Existing primary studies frequently fail to distinguish between these competing explanations. Third, healthcare quality improvement

evaluations often inadequately account for secular trends, temporal changes, and the sustainability of intervention effects, potentially introducing bias into conventional before-and-after comparisons (Dixon-Woods & Martin, 2016; Reed & Card, 2016; Knudsen et al., 2019).

To address these gaps, this study introduces four methodological contributions. First, the Reporting Yield Decomposition framework separates observed incident reporting into three components: exposure, the true incident rate, and reporting propensity. A severity-stratified estimator, termed the Reporting Propensity Index (RPI), is proposed to estimate changes in reporting propensity under specified identification assumptions. Second, partial identification bounds are developed to account for violations of these assumptions and quantify the resulting uncertainty. Third, a behavioral reporting model is formulated to identify the organizational and contextual conditions under which Kaizen-based interventions may be most effective, providing a testable hypothesis regarding potential moderators of intervention outcomes. Fourth, a sustainability half-life measure is introduced to relate changes in interrupted time series (ITS) trends to the durability of intervention effects. Collectively, these components provide a framework for interpreting changes in incident reporting beyond conventional report counts. Four Monte Carlo simulation studies were conducted to evaluate the statistical properties of the proposed methods, and a systematic review protocol was developed to guide their subsequent application.

Accordingly, this study aimed to develop and validate an analytical framework capable of distinguishing changes in reporting propensity from changes in actual incident occurrence following Kaizen-based patient safety interventions and to establish a systematic review protocol for its application. Theoretically, the study contributes to patient safety research by providing a formal measurement framework for addressing the reporting paradox and strengthening inferences about organizational reporting culture. Practically, the proposed framework may assist hospital managers and patient safety teams in evaluating Kaizen-based improvement programs by integrating the RPI with severity-stratified incident rates. This approach is intended to support more informed decisions regarding intervention effectiveness, distinguish improvements in reporting practices from potential increases in patient harm, and guide sustainable patient safety improvement strategies.

METHOD

Conceptual and Mathematical Framework

The proposed framework was developed to distinguish the effects of Kaizen-based interventions on incident reporting behavior from their effects on actual patient harm. It comprised three components: decomposition of the observed reporting rate, estimation of reporting propensity using identification assumptions and partial identification bounds, and a behavioral and dynamic model designed to examine heterogeneity in intervention effects and their sustainability over time.

Reporting Yield Decomposition

Let unit i (ward or hospital) in period t accumulate exposure N (patient-days). Incidents of severity class s occur at true rate θ, and each is reported with probability π. Class L covers near misses and no-harm incidents; class H covers incidents causing moderate or greater harm. Report counts follow

$$R_{ist} \sim \text{Poisson}(N_{it}\,\theta_{ist}\,\pi_{ist}), \qquad \lambda_{ist} = \frac{\mathbb{E}[R_{ist}]}{N_{it}} = \theta_{ist}\,\pi_{ist}. \qquad (1)$$

For a Kaizen programme, the counterfactual-adjusted change in the log reporting rate of class *s* is

$$\Delta_s \equiv \ln\frac{\lambda_s^{K=1}}{\lambda_s^{K=0}} = \underbrace{\Delta\ln\theta_s}_{\text{true incidents}} + \underbrace{\Delta\ln\pi_s}_{\text{reporting propensity}} . \qquad (2)$$

This identity is the formal core of the reporting paradox: Δ, which primary studies report, is the sum of two effects of opposite desirability, and neither is identified from Δ alone.

Identification and Partial Identification

Assumption A1 (severity completeness). Reporting propensity for harmful incidents is stable: $\Delta \ln \pi_H = 0$. This is plausible where moderate and severe harm is captured through mandatory channels, mortality review or record audit.

Assumption A2 (severity-mix invariance). The programme changes true incidents proportionally across classes: $\Delta \ln \theta_L = \Delta \ln \theta_H$.

Proposition 1 (identification). Under A1 and A2, the Reporting Propensity Index satisfies

$$\text{RPI} \equiv \Delta_L - \Delta_H = \Delta\ln\pi_L, \qquad \Delta\ln\theta = \Delta_H. \qquad (3)$$

Proof. Subtract the decomposition for *H* from that for *L*: $\Delta_L - \Delta_H = (\Delta \ln \theta_L - \Delta \ln \theta_H) + (\Delta \ln \pi_L - \Delta \ln \pi_H)$. A2 sets the first bracket to zero and A1 sets $\Delta \ln \pi_H$ to zero. Substituting A1 into the decomposition for *H* gives $\Delta_H = \Delta \ln \theta_H = \Delta \ln \theta$. ■

A2 is strong, because Kaizen may prevent some kinds of incident more than others. Relax it to $|\Delta \ln \theta_L - \Delta \ln \theta_H| \le \delta$, and relax A1 to $0 \le \Delta \ln \pi_H \le \eta$, since a programme can only raise, not lower, already near-complete reporting of harm.

Proposition 2 (bounds and breakdown point). Under the relaxed assumptions,

$$\Delta\ln\pi_L \in [\,\text{RPI} - \delta,\ \text{RPI} + \delta + \eta\,], \qquad \delta^* = \text{RPI}\ \ (\eta = 0). \qquad (4)$$

Proof. From the subtraction above, $\Delta \ln \pi_L = \text{RPI} - (\Delta \ln \theta_L - \Delta \ln \theta_H) + \Delta \ln \pi_H$. The middle term lies in $[-\delta, \delta]$ and the last in $[0, \eta]$; the bounds follow. The conclusion "reporting propensity rose" survives as long as the lower bound is positive, i.e. $\delta < \text{RPI}$. ■

The breakdown point δ* plays the role of an E-value: it states how large a shift in the true severity mix, on the log scale, would be needed to overturn the finding. Because Δ_L and Δ_H come from independent Poisson strata, the sampling variance of the RPI for a before–after contrast is approximately

$$\widehat{\text{Var}}\big(\widehat{\text{RPI}}\big) \approx \sum_{s\in\{L,H\}} \left(\frac{1}{R_s^{\text{pre}}} + \frac{1}{R_s^{\text{post}}}\right)\cdot\hat{\phi}, \qquad (5)$$

where $\hat{\varphi}$ is the overdispersion factor ($\hat{\varphi} = 1$ under pure Poisson variation). For ITS studies, the RPI is the difference of the stratum-specific level-change coefficients, with its variance taken from the joint model.

Table 1 turns the two identified quantities into a decision matrix for hospital managers.

Table 1. Interpretation of Kaizen effects by sign of the true-harm change and of the RPI

	RPI > 0 (reporting culture strengthened)	**RPI ≤ 0 (reporting culture unchanged or weakened)**
$\Delta_H < 0$ (harm fell)	Target state: safer care and more learning	Possible silencing; verify harm with record review
$\Delta_H \geq 0$ (harm rose or flat)	Better visibility; prioritise action on reported themes	Programme failure; redesign before scaling

Worked illustration (hypothetical values, for exposition only). Suppose no-harm reports rise from 8.0 to 12.0 per 1,000 patient-days, so $\Delta_L = \ln 1.5 = 0.405$, while harm reports fall from 1.0 to 0.9, so $\Delta_H = \ln 0.9 = -0.105$. Then RPI = 0.511: reporting propensity rose by a factor of $\exp(0.511) \approx 1.67$, and true harm fell by about 10%. The breakdown point $\delta^* = 0.511$ means the finding fails only if the programme shifted the true severity mix by a factor of 1.67 or more.

Behavioural Reporting Model

A staff member who observes an incident reports it when the expected benefit exceeds the cost. Let *b* be the perceived value of a fix, φ the probability that a report leads to visible action (feedback closure), and *c* the cost of reporting (time, effort, fear of blame). With a logistic error,

$$\pi = \Lambda(x) = \frac{1}{1 + e^{-x}}, \qquad x = b\,\varphi - c. \qquad (6)$$

Kaizen acts on both levers: huddles and visual boards raise φ, while simplified forms, 5S and a blame-free stance lower *c* (Mazzocato et al., 2016; Archer et al., 2017).

Proposition 3 (baseline dependence). For a programme that shifts the index by $\Delta x = b\Delta\varphi + \Delta c$ (Δc denoting the cost reduction),

$$\Delta \ln \pi \approx (1 - \pi_0)\,\Delta x, \qquad \Delta\pi \approx \pi_0(1 - \pi_0)\,\Delta x. \qquad (7)$$

Proof. $d \ln \Lambda(x)/dx = 1 - \Lambda(x)$ and $d\Lambda(x)/dx = \Lambda(x)(1 - \Lambda(x))$; a first-order Taylor expansion around x_0 gives both results. ■

Two testable predictions follow. **H1:** the relative effect (log rate ratio), which the meta-analysis pools, decreases with baseline reporting propensity, so settings with severe under-reporting gain most in relative terms. **H2:** because $\Delta\varphi$ enters multiplied by *b*, closing the feedback loop pays off most where staff believe reports lead to change, which links Kaizen effects to leadership and psychological safety (O'Donovan & McAuliffe, 2020; Grailey et al., 2021). Both are tested by meta-regression (see Synthesis Model).

Dynamics, Bias and Sustainability

For ITS data, the segmented Poisson model with exposure offset is

$$\ln \mathbb{E}[R_t] = \ln N_t + \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 (t - T_0) D_t + \sum_m \gamma_m S_{mt}, \qquad (8)$$

where $D_t = 1$ after implementation at T_0 and S are seasonal terms (Bernal et al., 2017). The effect *k* periods after implementation is $RR(k) = \exp(\beta_2 + \beta_3 k)$.

Proposition 4 (bias of naive before–after comparisons). With *m* pre-periods and *m* post-periods, the naive contrast of mean log rates equals the true mean effect plus $\beta_1 m$.

Proof. Mean pre-period predictor: $\beta_0 + \beta_1(m+1)/2$. Mean post-period predictor: $\beta_0 + \beta_1(3m+1)/2 + \beta_2 + \beta_3(m+1)/2$. Their difference is $\beta_1 m + [\beta_2 + \beta_3(m+1)/2]$, and the bracket is the true mean post-period effect. ■

Because accreditation and national campaigns push reporting upward over time ($\beta_1 > 0$), uncontrolled studies overstate Kaizen effects, and the bias grows with window length. This motivates the design restrictions in the eligibility criteria.

Finally, many improvement programmes show an early jump that fades (Hallam & Contreras, 2018; Leite et al., 2020). Let the log effect decay as $\ln RR(k) = \alpha\, e^{-\kappa k}$. The **sustainability half-life**

$$h = \frac{\ln 2}{\kappa} \qquad (9)$$

is the number of periods after which half of the initial effect remains. It is estimated from studies with at least twelve post-implementation points and summarises durability in a single, comparable number.

Systematic Review Protocol

Protocol and Reporting

The review follows PRISMA 2020 (Page et al., 2021) and, where pooling is not possible, the SWiM guideline (Campbell et al., 2020). The protocol was registered in PROSPERO before screening [registration ID]. Deviations from the protocol are reported in Supplementary File S1.

Eligibility Criteria

Table 2. Eligibility criteria (PICOS)

Element	Inclusion	Exclusion
Population	Hospitals, wards, primary care and long-term care units; any staff group	Simulation-only or student-only settings
Intervention	Programme with an explicit Kaizen component: daily kaizen, kaizen events (rapid improvement events), kaizen huddles or boards, 5S–Kaizen–TQM, PDCA-based continuous improvement within Lean	Lean or Six Sigma without a Kaizen element; single-component IT reporting upgrades
Comparator	Usual practice, pre-implementation period, or non-implementing units	No comparator and fewer than three pre-points
Outcomes	Primary: incident reporting rate per exposure. Secondary: reports by severity class, near-miss share, feedback time, safety culture scores (HSOPSC, SAQ), harm rate	Outcomes without exposure or denominator that cannot be recovered
Study design	RCT, cluster RCT, controlled before–after (CBA), ITS with at least three points per phase	Uncontrolled before–after, case reports, editorials
Time and language	Published January 2016 – [search date] 2026; English or Indonesian	—

The design restrictions follow directly from Proposition 4: uncontrolled before–after studies carry a bias of $\beta_1 m$ that cannot be estimated, so they are excluded from quantitative synthesis and summarised descriptively only.

Information Sources and Search

MEDLINE (PubMed), Embase, Scopus, Web of Science Core Collection, CINAHL and Garuda/SINTA were searched, supplemented by citation chasing of included studies and relevant reviews (Moraros et al., 2016; Tlapa et al., 2020; Marsilio & Pisarra, 2021). The core Scopus string was:

TITLE-ABS-KEY((kaizen OR "continuous improvement" OR "rapid improvement event*" OR "5S" OR "lean management" OR "lean healthcare" OR "improvement huddle*" OR PDCA) AND ("incident report*" OR "adverse event report*" OR "near miss*" OR "safety report*" OR "error report*" OR "patient safety") AND (hospital* OR ward* OR clinic* OR "health care" OR healthcare)) AND PUBYEAR > 2015

Strings were translated to each database's syntax and are listed in Supplementary File S2.

Selection and Data Extraction

Two reviewers independently screened titles, abstracts and full texts in Rayyan (Ouzzani et al., 2016); disagreements were resolved by a third reviewer, and agreement is reported as Cohen's κ. Extraction used a piloted form covering setting, country income group, Kaizen components, implementation intensity, duration, leadership involvement, baseline reporting rate, counts and exposure by severity class, ITS coefficients and follow-up length. Kaizen intensity was coded on a 0–4 scale (one point each for daily kaizen, structured events, visual management, and closed-loop feedback to reporters), following component-based definitions of Lean (Rotter et al., 2019).

Risk of Bias and Certainty

Randomised studies were appraised with RoB 2 (Sterne et al., 2019); CBA and ITS studies with ROBINS-I (Sterne et al., 2016), supplemented by ITS-specific items on autocorrelation, seasonality and the number of points (Lopez Bernal et al., 2018). Certainty of each outcome was rated with GRADE (Hultcrantz et al., 2017). For the RPI, the indirectness domain additionally considers plausibility of A1 and A2, judged against the breakdown point δ*.

Effect Measures

The effect measure is the log rate ratio (log RR) of reports per 1,000 patient-days. For CBA studies, $y = \ln(\lambda_{post}/\lambda_{pre})_K - \ln(\lambda_{post}/\lambda_{pre})_C$. For ITS studies, raw series were re-analysed where available with the segmented Poisson model of Equation (8), with Newey–West or ARMA errors for autocorrelation (Bernal et al., 2017); the primary ITS effect is ln RR at 12 months, $\beta_2 + 12\beta_3$. Severity-stratified estimates Δ_L and Δ_H were extracted or derived, and the RPI with its variance computed with Equations (3) and (5).

Synthesis Model

Effects were pooled with a random-effects model:

$$y_i = \mu + \mathbf{x}_i^{\top}\boldsymbol{\gamma} + u_i + \varepsilon_i, \qquad u_i \sim \mathcal{N}(0, \tau^2), \quad \varepsilon_i \sim \mathcal{N}(0, v_i). \qquad (10)$$

The between-study variance τ^2 was estimated by REML, which performs well across simulation scenarios (Veroniki et al., 2016), and confidence intervals used the Hartung–Knapp–Sidik–Jonkman adjustment. A 95% prediction interval, $\hat{\mu} \pm t(k-2)\sqrt{(\hat{\tau}^2 + SE(\hat{\mu})^2)}$, is reported for every pooled estimate because it describes the effect expected in a new hospital (IntHout et al., 2016). When counts were sparse (fewer than five events in any cell), a Poisson–normal generalised linear mixed model replaced the inverse-variance model (Bakbergenuly & Kulinskaya, 2018).

Hypotheses H1 and H2 were tested by meta-regression with covariates x_i: baseline reporting rate (reports per 1,000 patient-days, standardised), measured in a period that precedes and does not overlap the pre-intervention window of the effect estimate, to avoid regression to the mean (see Study C in the Results), Kaizen intensity score, leadership

involvement, and country income group. H1 predicts a negative coefficient on baseline propensity; H2 predicts a positive interaction between closed-loop feedback and leadership involvement. With fewer than ten studies per covariate, meta-regression is reported as exploratory.

The sustainability half-life was estimated by fitting $\ln RR(k) = \alpha\, e^{-\kappa k}$ to the post-implementation effects of each ITS with at least twelve post-points, then pooling $\ln \kappa$ in a random-effects model. Heterogeneity is reported as τ^2, I^2 and prediction intervals; small-study effects are examined with funnel plots and Egger's test when at least ten studies are available.

Sensitivity Analyses

Pre-specified sensitivity analyses: (a) excluding studies at serious or critical risk of bias; (b) using DerSimonian–Laird instead of REML; (c) restricting to studies with independent harm ascertainment (record review), where A1 is most credible; (d) RPI bounds at $\delta = 0.1$, 0.2 and 0.3 and $\eta = 0.1$; and (e) leave-one-out influence analysis.

Monte Carlo Evaluation

Before applying the framework to primary studies, its estimators were tested in four simulation studies designed and reported with the ADEMP structure: aims, data-generating mechanisms, estimands, methods and performance measures (Morris et al., 2019). Table 3 summarises the design.

Table 3. Design of the simulation studies

Study	Aim	Data-generating mechanism	Estimand and methods compared	Performance measures
A (Propositions 1–2)	Recover the change in reporting propensity from report counts	k = 15 studies; exposure 20,000–60,000 patient-days per period; true rates 10 (class L) and 1 (class H) per 1,000 patient-days; $\pi_H = 0.90$; baseline $\pi_L \in \{0.2, 0.5\}$ (study SD 0.3 on logit scale); Kaizen shift $\Delta x = 0.8$ ($\tau = 0.15$); true harm change $g \in \{-0.2, 0, +0.1\}$; severity-mix violation $\delta \in \{0, 0.1, 0.3\}$; harm-propensity rise $\eta \in \{0, 0.05\}$; gamma overdispersion (shape 50); 36 scenarios × 1,000 replications	Mean $\Delta \ln \pi_L$; pooled RPI vs pooled total-report log RR, both REML with HKSJ	Bias, RMSE, 95% CI coverage, Table 1 classification accuracy, coverage of bounds at $\delta = 0.3$, $\eta = 0.1$
B (Proposition 4)	Quantify bias of uncontrolled before–after designs	Monthly series, 3,000 patient-days and 10 reports per 1,000 per month; $\beta_2 = 0.30$, $\beta_3 = 0$; pre-trend $\beta_1 \in \{0, 0.005, 0.010\}$; $m \in \{12, 24\}$; 1,000 replications each	Mean post-period effect; naive difference of mean log rates vs segmented quasi-Poisson ITS	Bias against the theoretical $\beta_1 m$; ITS CI coverage

Study	Aim	Data-generating mechanism	Estimand and methods compared	Performance measures
C (Proposition 3, H1)	Test whether meta-regression detects baseline dependence	$k \in \{15, 30\}$; baseline π_L uniform on the logit scale over 0.1–0.7; H1 true (logistic shift $\Delta x = 0.8$) or false (constant log gain 0.40); covariate from the same baseline period or an independent earlier period; 2,000 replications	Slope on standardised log baseline rate; DerSimonian–Laird τ^2 with Knapp–Hartung test	Mean slope; rejection rate ($\alpha = 0.05$, negative slope)
D (half-life)	Recover the sustainability half-life	24 pre-months, 24 or 36 post-months; pre-trend 0.003; ln RR(k) = $0.5e^{-0.05k}$, so h = 13.9 months; 1,000 replications; pooled over 5, 10 or 20 studies	h from counterfactual-based non-linear least squares; pooling on log κ	Median, IQR and 95% range; failure rate

Simulations were run in Python 3 (NumPy 2.4, SciPy 1.17, statsmodels 0.15) with fixed seeds; code is in Supplementary File S6.

RESULTS AND DISCUSSION

Results of the Monte Carlo Evaluation

The RPI recovered the change in reporting propensity without bias when its assumptions held, while the conventional total-count estimator was biased by up to −0.35 on the log scale. The sections below report the four simulation studies. The empirical synthesis of primary studies is being conducted under the registered protocol described in the Method section and will be reported in a companion paper.

Study A: Identification of Reporting Propensity

Under assumptions A1 and A2, the true mean change in log reporting propensity was about 0.58 at baseline $\pi_L = 0.2$ and 0.32 at $\pi_L = 0.5$, already consistent with Proposition 3. Table 4 compares the two estimators in these scenarios.

Table 4. Performance when A1 and A2 hold ($\delta = 0$, $\eta = 0$; k = 15; 1,000 replications per row)

Baseline π_L	True harm change g	RPI bias	RPI RMSE	RPI 95% CI coverage	Total-count bias	Total-count RMSE	Total-count coverage
0.2	−0.2	−0.003	0.107	0.957	−0.350	0.355	0.001
0.2	0	−0.001	0.105	0.948	−0.149	0.160	0.371
0.2	+0.1	+0.003	0.106	0.946	−0.048	0.075	0.902
0.5	−0.2	−0.002	0.101	0.956	−0.244	0.251	0.034
0.5	0	0.000	0.102	0.950	−0.045	0.072	0.901
0.5	+0.1	+0.002	0.101	0.951	+0.056	0.080	0.857

The total-count estimator failed most when Kaizen succeeded on both fronts: when true harm fell by 18% (g = −0.2), falling harm reports masked rising near-miss reports, and coverage dropped to 0.1–3.4%. This is the reporting paradox in numerical form.

When assumptions were violated, the RPI's bias equalled $\delta - \eta$ almost exactly, as Proposition 2 implies (Table 5). Point-estimate coverage fell accordingly, but the partial-

identification interval, built from the RPI's 95% CI widened by $\delta = 0.3$ and $\eta = 0.1$, covered the truth in 97.4–100% of replications in every scenario.

Table 5. RPI performance under violated assumptions (averaged over baseline π_L and g; 6,000 replications per row)

Severity-mix violation δ	Harm-propensity rise η	Theoretical bias (δ − η)	Observed mean RPI bias	RPI CI coverage	Bound coverage
0	0	0	0.000	0.946–0.957	1.000
0	0.05	−0.05	−0.047	0.921–0.944	1.000
0.1	0	0.10	0.103	0.835–0.873	0.999–1.000
0.1	0.05	0.05	0.054	0.915–0.943	1.000
0.3	0	0.30	0.301	0.198–0.295	0.974–0.986
0.3	0.05	0.25	0.250	0.359–0.431	0.989–0.995

For decision-making, the Table 1 quadrant assigned from pooled Δ_H and RPI was correct in 95.2–99.2% of replications when harm truly fell. A reader using the total count alone was correct in none, because a single number cannot show falling harm and rising reporting at once. When harm truly rose ($g = +0.1$), both moved in the same direction; the total count was then correct in 99.9–100% and the RPI-based rule in 87.5–97.3%.

Study B: Bias of Uncontrolled Before–After Designs

The naive before–after estimate tracked the bias predicted by Proposition 4 closely, while the segmented ITS model removed it (Table 6).

Table 6. Bias of the post-period effect (true $\beta_2 = 0.30$; 1,000 replications per row)

Window m (months per phase)	Pre-trend β_1	Theoretical bias $\beta_1 m$	Naive bias	ITS bias	ITS 95% CI coverage
12	0	0	0.006	0.006	0.920
12	0.005	0.060	0.067	0.002	0.944
12	0.010	0.120	0.128	0.007	0.944
24	0	0	0.005	0.009	0.946
24	0.005	0.120	0.122	−0.003	0.953
24	0.010	0.240	0.246	−0.015	0.958

A modest secular rise of 1% per month, typical during accreditation cycles, inflated a two-year naive evaluation by 0.25 on the log scale, i.e. the reported effect was about 28% larger than the true effect.

Study C: Detecting Baseline Dependence (H1)

The moderator test was valid only when the covariate came from a period independent of the effect estimate (Table 7). When baseline reports served both as the covariate and as the denominator of the effect, shared sampling noise produced a spurious negative slope through regression to the mean. Under a true null this produced false-positive rates of 11.3% ($k = 15$) and 20.7% ($k = 30$), and the error grew with the number of studies.

Table 7. Meta-regression of the log rate ratio on standardised baseline reporting rate (2,000 replications per row)

Studies k	Covariate source	H1 true: mean slope	H1 true: power	H1 false: mean slope	H1 false: rejection rate
15	Same baseline period	−0.195	0.821	−0.056	0.113
30	Same baseline period	−0.199	0.992	−0.059	0.207
15	Independent earlier period	−0.144	0.493	−0.004	0.027
30	Independent earlier period	−0.145	0.818	−0.008	0.034

With an independent baseline, the one-sided rejection rate under the null stayed at the nominal 2.5–3.4%, and power reached 82% with 30 studies. The Synthesis Model and the analysis code (Supplementary File S4) were revised accordingly.

Study D: Sustainability Half-Life

Individual series estimated the half-life imprecisely. With 24 post-months, the median estimate was 11.3 months (IQR 5.4–26.9) against a true 13.9 months; with 36 post-months it was 10.6 months (IQR 5.0–27.9), and 6–8% of fits failed to converge to a decaying curve. Pooling on log κ narrowed the 95% range of the pooled estimate from 3.8–39.8 months (5 studies) to 5.3–27.0 (10 studies) and 6.6–21.0 (20 studies). The median stayed near 11.5 months, so the estimator understates durability by about 17%. The half-life is therefore conservative and should be reported only as a pooled quantity from at least ten series.

Discussion

Theoretical Contribution

The framework reframes the evaluation of Kaizen in patient safety. Earlier syntheses asked whether Lean improves outcomes in general (Moraros et al., 2016; Akmal et al., 2020). This article asks a narrower and harder question: whether Kaizen changes *the willingness to report*, as distinct from the occurrence of harm. The Reporting Yield Decomposition shows why this question cannot be answered from report counts alone, and Proposition 1 shows what additional structure, a severity split with stated assumptions, makes it answerable. The decomposition generalises Macrae's (2016) argument that reporting systems measure attention to risk rather than risk itself, and gives it a testable form.

The behavioural model links Kaizen mechanisms to the barrier literature. Fear, effort and the absence of feedback, the barriers most often reported (Archer et al., 2017; Hamed & Konstantinidis, 2022; Oweidat et al., 2023), map onto the cost term c and the feedback term φ. Kaizen differs from a one-off reporting campaign because it acts on φ through a daily, visible loop between a reported problem and its fix (Mazzocato et al., 2016). Proposition 3 then predicts that relative gains are largest where baseline reporting is weakest, which turns a common observation about context into a falsifiable claim.

Methodological Contribution

Three tools are reusable beyond Kaizen. The RPI with its breakdown point gives any reporting intervention, digital or cultural, an outcome that cannot be gamed by simply generating more low-value reports. Proposition 4 quantifies the bias of uncontrolled before–after designs as $\beta_1 m$, which supports critiques that many PDSA-based projects lack the rigour

to support causal claims (Reed & Card, 2016; Knudsen et al., 2019). The sustainability half-life responds to concerns that Lean gains fade once attention moves on (Hallam & Contreras, 2018; Leite et al., 2020), and makes durability comparable across studies with different follow-up lengths. The simulations add a fourth, less obvious lesson. Testing whether Kaizen works best where reporting was weakest invites regression to the mean, because a unit with a chance low baseline will appear to improve most. Reusing the baseline period as the covariate raised the false-positive rate from about 3% to 11% at k = 15 and to 21% at k = 30 (Study C). Any review that relates improvement to baseline performance, a common analysis in quality-improvement research, should measure the baseline in an independent window.

Practical Implications

Hospital managers should stop treating a rising report count as a success metric on its own. Reports should be classified by severity at entry, and the RPI and Δ_H tracked together, using Table 1 to decide whether to scale, verify or redesign a programme. Near-miss reports deserve priority because they carry the learning signal before harm occurs (Woodier et al., 2023); a rising near-miss share with stable or falling harm is the clearest sign that Kaizen is working.

For Indonesian hospitals, the model suggests specific design choices. Reporting is low and constrained by blame culture and unclear processes (Dhamanti et al., 2019, 2020). Proposition 3 implies large relative gains are possible, but only if the cost term is cut, for example through a one-page report form, protected time in daily huddles and an explicit non-punitive policy. Simple 5S–Kaizen–TQM approaches have been applied in resource-limited hospitals with modest investment (Kanamori et al., 2016), which makes them a feasible entry point. Accreditation by KARS and national patient-safety regulation will raise reporting trends regardless of Kaizen; evaluations must therefore model the secular trend, as Proposition 4 shows.

Kaizen also carries risks. Poorly led programmes can intensify work without adding resources (Rees & Gauld, 2017), and frontline effects vary with how much autonomy staff are given (Mahmoud et al., 2021). Psychological safety and leadership involvement condition whether staff believe a report will lead to change (O'Donovan & McAuliffe, 2020; Grailey et al., 2021), which is the parameter *b* in the model and the focus of hypothesis H2. A broader safety culture remains the background condition for any of these effects (Reis et al., 2018; Lee et al., 2019; Marsilio & Pisarra, 2021).

Limitations

First, the RPI depends on assumptions A1 and A2; the bounds in Proposition 2 and sensitivity analysis (d) make this dependence explicit but cannot remove it. Second, severity coding varies between reporting systems and over time, and recoding during a Kaizen programme would bias Δ_L and Δ_H. Third, the behavioural model is a reduced form: it does not represent team-level spill-overs or strategic under-reporting. Fourth, the evidence base is expected to include few randomised studies, so most findings will rest on ITS and CBA designs rated at moderate or low certainty. Finally, Kaizen is rarely delivered alone, and co-interventions such as accreditation, electronic reporting upgrades or leadership walkrounds may confound effects even in controlled designs (Dixon-Woods & Martin, 2016). The simulations used stylised data-generating mechanisms with two severity classes, gamma overdispersion and no autocorrelation in Study A; real reporting systems are messier, and

performance should be rechecked against the empirical data once the review is complete. Last, this paper reports the framework, its validation and the protocol; the pooled empirical estimates will follow in the companion review.

CONCLUSION

Kaizen-based programs offer a promising approach to strengthening patient safety incident reporting; however, their evaluation remains challenging because of the reporting paradox, whereby an increase in reported incidents may indicate either improved reporting practices or increased patient harm. This study addressed this challenge by developing and validating an analytical framework that distinguishes changes in reporting propensity from changes in actual incident occurrence. The Reporting Yield Decomposition framework separates exposure, the true incident rate, and reporting propensity, while the Reporting Propensity Index (RPI) estimates changes in reporting propensity under specified identification assumptions and provides partial identification bounds when these assumptions are violated. The behavioral reporting model predicts that relative improvements in reporting propensity are greatest in settings with substantial baseline underreporting, whereas the sustainability half-life measure characterizes the durability of intervention effects. Monte Carlo simulations demonstrated that the RPI estimator was unbiased when its assumptions were satisfied, while its partial identification bounds maintained the specified coverage under assumption violations. In contrast, the conventional total-count estimator frequently misinterpreted the effects of successful interventions. Future evaluations of Kaizen-based programs should incorporate severity-stratified incident counts with appropriate exposure denominators, employ interrupted time series (ITS) or controlled study designs with at least 12 postintervention observations, and report RPI estimates alongside their sensitivity bounds and breakdown points to support more reliable assessments of reporting culture and patient safety outcomes.

FUNDING

This study did not receive any specific funding from governmental institutions, private organizations, or other funding agencies.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to the conduct of the research, preparation, or publication of this article.

AUTHOR CONTRIBUTIONS

All authors contributed to the preparation of this article. Andan Suryo Wardani, Dyah Wiji Puspita Sari, Dwi Heppy Rochmawati and Muh. Abdurrouf, was responsible for the conceptualization of the study, data collection and analysis, interpretation of the results, and manuscript preparation. The other authors contributed to the review, editing, and final approval of the manuscript.

GENERATIVE AI STATEMENT

The authors declare that no generative artificial intelligence was used in the preparation of this article.

REFERENCES

Akmal, A., Greatbanks, R., & Foote, J. (2020). Lean thinking in healthcare – Findings from a systematic literature network and bibliometric analysis. *Health Policy*, *124*(6), 615–627. https://doi.org/10.1016/j.healthpol.2020.04.008

Archer, S., Hull, L., Soukup, T., Mayer, E., Athanasiou, T., Sevdalis, N., & Darzi, A. (2017). Development of a theoretical framework of factors affecting patient safety incident reporting: A theoretical review of the literature. *BMJ Open*, *7*(12), e017155. https://doi.org/10.1136/bmjopen-2017-017155

Bakbergenuly, I., & Kulinskaya, E. (2018). Meta-analysis of binary outcomes via generalized linear mixed models: A simulation study. *BMC Medical Research Methodology*, *18*, 70.

Bates, W., D., Levine, M., D., Salmasian, H., Syrowatka, A., Shahian, M., D., Lipsitz, S., Zebrowski, P., J., Myers, C., L., Logan, S., M., Roy, … … Mort, E. (2023). The safety of inpatient health care. *New England Journal of Medicine*, *388*(2), 142–153. https://doi.org/10.1056/NEJMsa2206117

Campbell, M., McKenzie, E., J., Sowden, A., Katikireddi, V., S., Brennan, E., S., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, … Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, *368*, l6890. https://doi.org/10.1136/bmj.l6890

Dhamanti, I., Leggat, S., & Barraclough, S. (2020). Practical and cultural barriers to reporting incidents among health workers in Indonesian public hospitals. *Journal of Multidisciplinary Healthcare*, *13*, 351–359. https://doi.org/10.2147/JMDH.S240124

Dhamanti, I., Leggat, S., Barraclough, S., & Tjahjono, B. (2019). Patient safety incident reporting in Indonesia: An analysis using World Health Organization characteristics for successful reporting. *Risk Management and Healthcare Policy*, *12*, 331–338. https://doi.org/10.2147/RMHP.S222262

Dixon-Woods, M., & Martin, G. P. (2016). *Does quality improvement improve quality? Future Hospital Journal, 3(3), 191–194*. https://doi.org/10.7861/futurehosp.3-3-191

Grailey, E., K., Murray, E., Reader, T., & Brett, S. J. (2021). The presence and potential impact of psychological safety in the healthcare setting: An evidence synthesis. *BMC Health Services Research*, *21*, 773. https://doi.org/10.1186/s12913-021-06740-6

Hallam, C. R. A., & Contreras, C. (2018). Lean healthcare: Scale, scope and sustainability. *International Journal of Health Care Quality Assurance*, *31*(7), 684–696. https://doi.org/10.1108/IJHCQA-02-2017-0023

Hamed, M. M. M., & Konstantinidis, S. (2022). Barriers to incident reporting among nurses: A qualitative systematic review. *Western Journal of Nursing Research*, *44*(5), 506–523. https://doi.org/10.1177/0193945921999449

Hultcrantz, M., Rind, D., Akl, A., E., Treweek, S., Mustafa, A., R., Iorio, A., Alper, S., B., Meerpohl, J., J., Murad, H., M., Ansari, … Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *Journal of Clinical Epidemiology*, *87*, 4–13. https://doi.org/10.1016/j.jclinepi.2017.05.006

Kanamori, S., Shibanuma, A., & Jimba, M. (2016). Applicability of the 5S management method for quality improvement in health-care facilities: A review. *Tropical Medicine and Health*, *44*, 21.

Knudsen, V., S., Laursen, B., H. V, Johnsen, P., S., Bartels, D., P., Ehlers, L. H., & Mainz, J. (2019). Can quality improvement improve the quality of care? A systematic review of

reported effects and methodological rigor in plan-do-study-act projects. *BMC Health Services Research*, *19*, 683. https://doi.org/10.1186/s12913-019-4482-6

Lee, E., S., Scott, D., L., Dahinten, S., V., Vincent, C., Lopez, K. D., & Park, C. G. (2019). Safety culture, patient safety, and quality of care outcomes: A literature review. *Western Journal of Nursing Research*, *41*(2), 279–304. https://doi.org/10.1177/0193945917747416

Leite, H., Bateman, N., & Radnor, Z. (2020). Beyond the ostensible: An exploration of barriers to lean implementation and sustainability in healthcare. *Production Planning & Control*, *31*(1), 1–18.

Macrae, C. (2016). The problem with incident reporting. *BMJ Quality & Safety*, *25*(2), 71–75. https://doi.org/10.1136/bmjqs-2015-004732

Mahmoud, Z., Angelé-Halgand, N., Churruca, K., Ellis, L. A., & Braithwaite, J. (2021). The impact of lean management on frontline healthcare professionals: A scoping review of the literature. *BMC Health Services Research*, *21*, 383. https://doi.org/10.1186/s12913-021-06344-0

Marsilio, M., & Pisarra, M. (2021). Lean management in health care: A review of reviews of socio-technical components for effective impact. *Journal of Health Organization and Management*, *35*(4), 475–491. https://doi.org/10.1108/JHOM-06-2020-0241

Mazzocato, P., Stenfors-Hayes, T., Schwarz, von T., U., Hasson, H., & Nyström, M. E. (2016). Kaizen practice in healthcare: A qualitative analysis of hospital employees' suggestions for improvement. *BMJ Open*, *6*(7), e012256. https://doi.org/10.1136/bmjopen-2016-012256

Mitchell, I., Schuster, A., Smith, K., Pronovost, P., & Wu, A. (2016). Patient safety incident reporting: A qualitative study of thoughts and perceptions of experts 15 years after 'To Err is Human.' *BMJ Quality & Safety*, *25*(2), 92–99.

Moraros, J., Lemstra, M., & Nwankwo, C. (2016). Lean interventions in healthcare: Do they actually work? A systematic literature review. *International Journal for Quality in Health Care*, *28*(2), 150–165. https://doi.org/10.1093/intqhc/mzv123

Morris, P., T., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

O'Donovan, R., & McAuliffe, E. (2020). A systematic review of factors that enable psychological safety in healthcare teams. *International Journal for Quality in Health Care*, *32*(4), 240–250. https://doi.org/10.1093/intqhc/mzaa025

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n7.

Panagioti, M., Khan, K., Keers, N., R., Abuzour, A., Phipps, D., Kontopantelis, E., Bower, P., Campbell, S., Haneef, R., Avery, A. J., & Ashcroft, D. M. (2019). Prevalence, severity, and nature of preventable patient harm across medical care settings: Systematic review and meta-analysis. *BMJ*, *366*, l4185. https://doi.org/10.1136/bmj.l4185

Reed, J. E., & Card, A. J. (2016). The problem with Plan-Do-Study-Act cycles. *BMJ Quality & Safety*, *25*(3), 147–152. https://doi.org/10.1136/bmjqs-2015-005076

Rees, G. H., & Gauld, R. (2017). *Can lean contribute to work intensification in healthcare? Journal of Health Organization and Management, 31(3), 369–384*. https://doi.org/10.1108/JHOM-11-2016-0219

Reis, T., C., Paiva, S. G., & Sousa, P. (2018). The patient safety culture: A systematic review by characteristics of Hospital Survey on Patient Safety Culture dimensions. *International Journal for Quality in Health Care*, *30*(9), 660–677. https://doi.org/10.1093/intqhc/mzy080

Rotter, T., Plishka, C., Lawal, A., Harrison, L., Sari, N., Goodridge, D., Flynn, R., Chan, J., Fiander, M., Poksinska, … Kinsman, L. (2019). What is lean management in health care? Development of an operational definition for a Cochrane systematic review. *Evaluation & the Health Professions*, *42*(3), 366–390. https://doi.org/10.1177/0163278718756992

Sterne, C., J. A., Hernán, A., M., Reeves, C., B., Savović, J., Berkman, D., N., Viswanathan, M., Henry, D., Altman, G., D., Ansari, T., M., Boutron, … … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, C., J. A., Savović, J., Page, J., M., Elbers, G., R., Blencowe, S., N., Boutron, I., Cates, J., C., Cheng, H.-Y., Corbett, S., M., Eldridge, … … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Tlapa, D., Zepeda-Lugo, A., C., Tortorella, L., G., Baez-Lopez, A., Y., Limon-Romero, J., Alvarado-Iniesta, A., & Rodriguez-Borbon, M. I. (2020). Effects of lean healthcare on patient flow: A systematic review. *Value in Health*, *23*(2), 260–273. https://doi.org/10.1016/j.jval.2019.11.002

Woodier, N., Burnett, C., & Moppett, I. (2023). The value of learning from near misses to improve patient safety: A scoping review. *Journal of Patient Safety*, *19*(1), 42–47. https://doi.org/10.1097/PTS.0000000000001078