SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Reducing Not-In-Good-Order Rates in Insurance and Structured Settlement Applications: A Simulated Lean Six Sigma Intervention with Statistical Process Control and Modelled Cycle-Time and Rework Outcomes

Krishna Deep Valluru

Senior Business Process Engineer Independent Researcher, India

Abstract

The title's reference to measured outcomes means measured on the synthetic population described in Section III. With that stated first, the substance is this. A not-in-good-order application is one that cannot proceed through normal processing because required information, signatures, documents, forms or eligibility details are missing, incorrect or inconsistent, and must leave the standard workflow for exception handling. The operational literature on such defects in insurance and structured settlement intake is, so far as a systematic search can establish, empty: the entire evidence base is vendor and trade material, with no refereed treatment at all. This paper specifies and exercises an analysis protocol for that problem end to end, on a synthetic population of 3,700 applications across twelve monthly subgroups. The protocol has a deliberate ordering. The operational definition is repaired first, because inconsistent reviewer classification of borderline defects is a measurement system defect rather than a process defect, and a baseline measured through an unreliable classifier cannot anchor anything that follows. Only then are process changes introduced, in four stages: capture redesign, validation at point of entry, standardised training and decision rules, and upstream automation of the controls that proved stable. Monitoring pairs a p-chart with control limits recalculated for each monthly subgroup, since a monthly census varies between roughly 280 and 330 applications, with an individuals and moving range chart on monthly mean cycle time. The paper's substantive argument concerns a root cause that could not be eliminated. Suitability and replacement exceptions depend on regulatory requirements, product rules and approvals outside the operating team's control, and their incidence does not move across the two periods. The response is to stop attacking the defect and attack its downstream cost instead, which is a distinction the improvement literature, framed almost uniformly around elimination, does not treat. What this paper can claim is that the protocol detects a shift of known magnitude under stated conditions and is internally coherent. What it cannot claim, and does not, is any effect in any real operation.

Keywords: not-in-good-order, Lean Six Sigma, DMAIC, statistical process control, p-chart, individuals chart, attribute measurement system analysis, operational definition, insurance operations, synthetic data, simulation study, method demonstration.

I. Introduction

An application that cannot proceed is not a rare event in insurance intake, and it is not well studied.

What the defect is, operationally. A not-in-good-order application is one that cannot move through normal processing because required information, signatures, documents, forms or eligibility details are missing, incorrect or inconsistent. It leaves the standard workflow and enters exception handling. In the synthetic population modelled here, one such event adds roughly three to four days of cycle time, about an hour of rework, and one or more additional handling touches. For the applicant it produces delay, repeated requests for information, and uncertainty.

The literature is absent, and that is a finding in itself. A systematic search returns no refereed treatment of not-in-good-order defects as an operations problem; queries surface unrelated insurance-law material. The commonly circulated incidence figures come from vendor and trade publications rather than from research. The closest published analogue anywhere is a Lean Six Sigma reduction of rejects in a retail bank's account-opening process [4], which is an adjacent intake-defect problem in financial services, and the insurance work that exists addresses claims cycle time [2] or framework-level instrument selection inside a carrier [3] rather than new-business intake.

Two consequences follow, and they cut in opposite directions. The absence is the paper's strongest motivation: a defect class that is operationally significant and widely discussed has no refereed treatment. It also means there is no published baseline incidence to calibrate against, so the synthetic baseline in Section III is a modelling assumption and is declared as one rather than presented as background.

The problem the paper actually attacks is upstream of the defect rate. Before the intervention, the classification was binary at first review, and different reviewers interpreted borderline cases differently. That is not a process problem. **It is a measurement system problem**, and it means a baseline computed from that classifier is unreliable in a way no amount of process improvement will correct. Attribute agreement analysis has been used to expose exactly this failure in manufacturing inspection [12] and the formal framework for binary measurement systems is established [11]; so far as this search can establish, neither has been applied to the classification of insurance application completeness. **The contribution here is a domain transfer, not an origination**, and Section IV states it that way.

Contributions.
1. **A protocol with measurement system repair sequenced first** (Sections IV and V), before any process change, so that the baseline the rest of the analysis rests on is produced by a classifier whose behaviour is known.
2. **A paired monitoring design** (Sections VI and VII): a p-chart with limits recalculated per subgroup, because the monthly census varies, alongside an individuals and moving range chart on mean cycle time. **This is textbook practice rather than a contribution**, and it is presented as coherence evidence: two responses shifting in the same month is weak evidence against the two having moved for unrelated reasons.
3. **The irreducible-cause treatment** (Section VIII), in which a root cause that provably does not move is managed by reducing its downstream cost rather than its incidence. **This is the paper's primary contribution**, and the improvement literature, framed around elimination, does not address it.
4. **A reporting-standards-conformant method demonstration on declared synthetic data** (Sections III and X), in a field with no tradition of it, in a domain where real data is almost never releasable.
Section II positions the work. Section III specifies the data-generating mechanism. Sections IV to VII give the protocol. Sections VII and VIII give results and the irreducible cause. Sections X and XI give limitations and conclude.

II. Related work

Lean Six Sigma in insurance operations exists and is close enough to require care. Sarkar, Mukhopadhyay and Ghosh applied DMAIC to insurance claim-processing cycle time and reported measurable reduction [2]. Sandner et al. gave a framework-level account of Lean and Six Sigma instrument selection inside a large insurer, with lessons drawn from a real multi-year deployment [3]. Sunder M reports reject reduction in a retail bank's account-opening process with quantified benefit [4], which is the closest published analogue to intake-defect reduction in financial services anywhere in the refereed record.
What this paper must not do is compete with those on outcome. Each reports real organisational results. This one reports none. **The differentiation is on sequencing and reproducibility:** measurement system repair before process change, which none of the three performs as an explicit first step; the irreducible-cause treatment, which none addresses; and a fully specified generating mechanism, which action research on proprietary data cannot offer. Those are procedural claims, and the paper makes them as procedural claims from the opening rather than arriving at them in discussion.
Statistical process control supplies the monitoring apparatus and is not in dispute. Shewhart's distinction between common and assignable causes and the economic rationale for three-sigma limits [5] is what licenses treating a post-intervention shift as a signal at all. Montgomery is the methodological reference for p-charts with variable subgroup size, individuals and moving range charts, Pareto analysis and first-pass yield [6], and Wheeler and Chambers supply the interpretation discipline for individuals charts, including why that chart is correct for one value per period and what a short baseline does and does not support [7]. The current international standard for Shewhart charts and their supplementary rules is ISO 7870-2:2023 [8], and Nelson's eight tests are the canonical formalisation of run rules [9].
One SPC issue is load-bearing and is addressed rather than left for a reviewer. At monthly subgroups of roughly 280 to 330, binomial p-charts are commonly overdispersed and generate false signals; Laney's p-prime chart is the standard correction [10]. Section VII reports the overdispersion check rather than assuming it away.
Measurement system analysis for attribute data is the backbone of the first phase. Van Wieringen and de Mast give the formal framework for binary classifications with capability indices analogous to gauge studies for continuous data [11]. Marques et al. demonstrate empirically that operator disagreement on borderline defect classification is itself a measurable system defect, that attribute agreement analysis quantifies it, and that definition plus training repairs it [12]. **That is precisely the argument this paper transfers to insurance intake**, and the transfer is the claim, not the idea.
Inference. Welch's unequal-variance test [13] is required rather than optional, because reducing variation is part of the intervention's purpose and the two periods' variances therefore differ by design. For proportions at

the counts involved, Agresti and Caffo's adjusted interval outperforms the Wald form [14], which matters in Section VIII where expected monthly counts in a single category are small.

Mistake-proofing and sustaining. Saurin, Ribeiro and Vidor give a structured framework for designing and assessing poka-yoke devices, including classification by where in the process the device acts [15], which supports point-of-entry validation with something more than an anecdote. McLean, Antony and Dahlgaard's systematic review of why continuous-improvement initiatives fail to hold [16] is the empirical justification for the control plan in Section IX and keeps that section from being a list of assertions.

Synthetic and simulation methodology, which is where this paper's framing comes from. Morris, White and Crowther's ADEMP structure requires the data-generating mechanism to be a first-class methods component rather than a footnote [17]. The STRESS guidelines make disclosure of model inputs, data sources and software mandatory [18]. Siepe et al. give a standardised template for planning and reporting methodological simulation work, including where transparency statements belong [19]. Raghunathan reviews synthetic data generation and, more importantly here, its inferential limits [20]. **Section III declares conformance with the first two by name.**

The gap this paper occupies. No Lean Six Sigma service-process study found declares a synthetic dataset. The improvement literature has no tradition of method demonstration on constructed data, whereas statistics and simulation research have mature conventions for exactly that. In a domain where real operational data is almost never releasable, a reporting-standards-conformant, fully reproducible demonstration is a legitimate form for the work to take, and it is the form adopted here.

III. The data-generating mechanism

This section is the data. It is placed before any analysis, per ADEMP [17] and the STRESS reporting guidelines [18], and the paper declares conformance with both.

Population and periods. Twelve monthly subgroups covering a single notional year. A baseline period of six months and a post-intervention period of six months, with the modelled intervention taking effect at the start of month seven.

TABLE I. The synthetic population, by period

	Baseline, months 1 to 6	**Post-intervention, months 7 to 12**
Applications	1,825	1,875
Not-in-good-order at first review	680	339
Rate	37.3 per cent	18.1 per cent
Monthly subgroup size	300 to 310	305 to 330
Mean cycle time	5.42 days	3.25 days
Total rework	1,197.9 hours	500.3 hours
Mean rework per application	39.4 minutes	16.0 minutes
Total analytical population	**3,700 applications**	

Subgrouping. Each monthly subgroup is a census of all eligible applications reviewed that month, not a sample. Subgroup size varies because volume varies, which is why the control limits in Section VII are recalculated for every subgroup rather than fixed.

The injected shift. The modelled intervention takes effect in month seven. Month seven carries 330 applications at a 21.2 per cent rate and a mean cycle time of 3.64 days. Subsequent months continue to decline modestly. **The magnitude of the shift is an input parameter, not a finding**, and Section VII says so where the result is reported.

The irreducible category. Suitability and replacement exceptions are generated at a fixed incidence of approximately 3.9 per cent of volume in both periods, 72 of 1,825 and 73 of 1,875. **This is a chosen input.** Its flatness across the two periods is a property of the generating mechanism, specified in order to exercise the protocol's handling of a cause that does not respond to the intervention. Section VIII treats the argument about such causes as the contribution and the number as a parameter.

Defect categories. Baseline defects are distributed across nine categories summing to the 680 baseline events, as given in Figure 3. The three largest account for 63.5 per cent and the five largest for approximately 84.6 per cent.

Monthly detail. The period aggregates above are the specification. The month-level series used to draw Figures 1 and 2 is reconstructed to reconcile exactly to every aggregate in Table I, and the reconstruction is checked programmatically: baseline and post application counts, defect counts, rates, the month-seven values and both period cycle-time means must all agree, or the figures do not build. **The monthly values are therefore a consistent realisation of the specification rather than additional data**, and the figure captions say so.

What is not modelled. No seasonality beyond the injected shift. No autocorrelation between months. No explicit reviewer-level model in the generated defect classifications; Section IV's measurement system argument is developed as protocol specification rather than exercised against simulated reviewer disagreement, and Section X records that as a limitation.

IV. The operational definition as the first defect

The classification was the problem before the process was. Before the intervention, a not-in-good-order determination was binary at first review: the application either could or could not proceed because of missing, incorrect, inconsistent or incomplete information, documentation, signatures, forms or required approvals. The difficulty is that reviewers interpreted borderline cases differently, which produces inconsistent classification and therefore unreliable baseline reporting.

Why this must be repaired first rather than alongside. A baseline rate is an estimate produced by a measuring instrument, and here the instrument is a human reviewer applying a definition. If two reviewers assign different outcomes to the same application, the rate is partly a function of who happened to review. Improving the process then changes a number whose movement cannot be cleanly attributed, because part of it is measurement variation. **Sequencing measurement system repair ahead of process change is the single most consequential ordering decision in the protocol**, and it is what the published intake-defect work does not do.

What the repair consisted of. Standardizing the definition, establishing a common defect taxonomy, and clarifying the counting rule, namely when an application is to be counted as not-in-good-order at all. The training that followed used **worked examples of borderline cases** so that reviewers would classify the same defect the same way. The author's framing of that step is exact and worth preserving. It addressed **both process variation and measurement system variation**.

What a full attribute study would require, and what this paper does not claim. A complete attribute measurement system analysis specifies the number of reviewers, the number of items, replicates per item, how borderline items were selected, the agreement statistic, and the threshold for acceptability [11], [12]. **The protocol specifies that such a study is run before the baseline is accepted. This paper does not report one**, because the generating mechanism in Section III does not model reviewer-level disagreement. Section X records this as the most significant gap between the protocol as specified and the demonstration as executed.

One design point that would otherwise bite. At a category incidence near 4 per cent, a fifty-item agreement study contains about two defective items, which cannot support a reliable agreement statistic. **The sample for an attribute study must therefore be deliberately stratified toward borderline cases**, and that stratification must be disclosed rather than left implicit.

V. The intervention, in four stages

The ordering matters, and each stage depends on the previous one being stable.

The protocol follows DMAIC as a standardised methodology rather than as a bespoke sequence, which matters because the phase deliverables are then defined externally, and the work is auditable against them. The international standard scopes DMAIC explicitly to service and transactional processes [1], which is the setting here. Sections IV and VI correspond to Define and Measure, this section to Analyse and Improve, and Section IX to Control. The one departure from the usual reading is the sequencing argument of Section IV: the measurement system is repaired within Measure, before Improve begins, rather than being assumed adequate.

First, capture redesign. The intake checklist was simplified, required fields clarified, ambiguous instructions removed, form versions standardised, and required supporting documents aligned to the specific application

type. The aim was to remove defects caused by unclear or inconsistent capture **before** they reach operations, which is the cheapest point at which to remove them.

Second, validation at point of entry. Required fields, signatures, dates, document attachments and key data-consistency checks were validated before a submission could move forward, with incomplete or contradictory entries flagged immediately so that the producer, customer or intake user could correct them before the application entered the processing queue. This is mistake-proofing in the sense formalised by Saurin et al. [15], and specifically the class of device that acts at the point of origin rather than detecting downstream.

Third, standardised training and work instructions. Intake staff and reviewers were trained on the revised definition, the defect taxonomy, the first-review decision rules and the escalation criteria, using borderline examples. This stage is simultaneously a process intervention and the measurement system repair of Section IV, and it is the stage that makes everything measured afterwards interpretable.

Fourth, upstream automation, deliberately last. Once the redesigned workflow and validation rules were stable, selected controls were moved upstream and automated: required-field checks, form-version validation, document-completeness checks, routing logic and exception alerts. **Automating earlier would have encoded rules that were still changing**, which is the common failure of this stage and the reason it is sequenced fourth rather than first.

VI. Monitoring design

Two responses, two charts, one intervention.

The attribute response. The monthly not-in-good-order rate at first review is a proportion over a census, so a p-chart is the correct instrument [6]. Because subgroup size varies, the control limits are recalculated for each subgroup rather than fixed. **The practical effect at these subgroup sizes is small**, since the range of roughly 280 to 330 is about nine per cent either side of the mean and the limits will look nearly flat. It is done because it is correct, not because it is visually consequential, and saying so is more honest than implying the variation does analytical work it does not do.

The continuous response. Monthly mean cycle time gives one value per period, which is the case an individuals and moving range chart exists for [6], [7]. The moving range measures change between consecutive monthly means and supplies the dispersion estimate that sets the individuals limits.

Run rules, declared rather than assumed. Applying all eight of Nelson's tests [9] to twelve points inflates the false-alarm rate well above nominal, and the supplementary rules in ISO 7870-2 [8] overlap with them without being identical. **The protocol declares a single rule set and the number of rules applied**, and the demonstration below relies on the most basic signal available, a point beyond a three-sigma limit, rather than on an accumulation of pattern rules.

VII. Results on the synthetic population

Every statement in this section describes the simulated process.



Figure 1. Monthly not-in-good-order rate on the synthetic population, drawn as a p-chart with control limits recalculated for each monthly subgroup.

The centre line is the baseline rate of 37.3 per cent. In the first month after the modelled intervention the simulated rate of 21.2 per cent on 330 applications falls below the applicable lower control limit of 29.3 per cent. The monthly series is a realisation reconciling exactly to the period aggregates in Table I.

The attribute signal. In the simulated process, month seven records 21.2 per cent against a baseline centre line of 37.3 per cent, with an applicable lower control limit of 29.3 per cent at that subgroup size. The point sits approximately six standard deviations below the centre line. Under the declared rule set this is a signal on the most conservative test available, and it does not depend on any pattern rule.

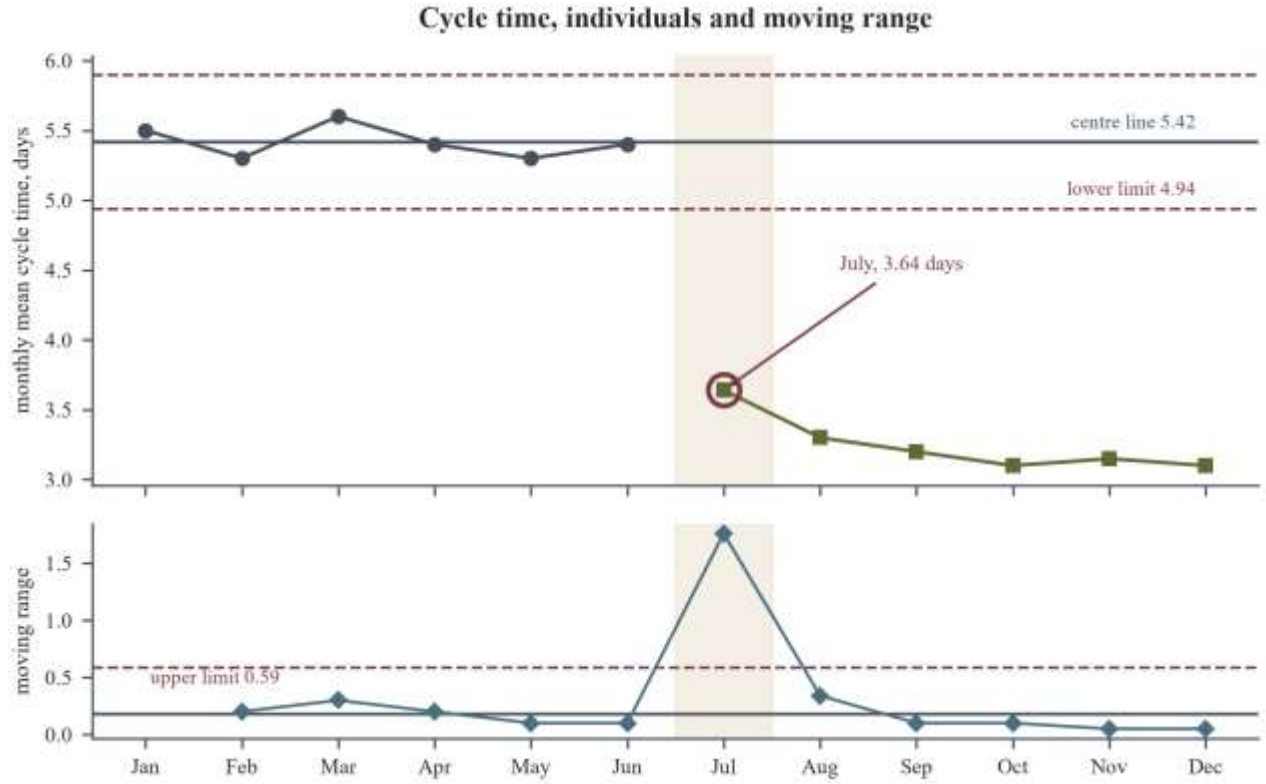


Figure 2. Monthly mean cycle time on the synthetic population, as an individuals and moving range chart. The individuals limits are derived from the baseline moving ranges. Month seven, at 3.64 days, falls below the baseline lower limit of 4.94 days, and the moving range for that month exceeds its upper limit, which is the expected signature of a step change rather than a drift.

The continuous signal, and what the moving range adds. Month seven's mean of 3.64 days falls below the baseline lower individuals limit. The corresponding moving range exceeds its own upper limit, which is informative in a way the individuals point alone is not: a step change produces exactly one large moving range at the step, whereas a gradual drift would produce a sequence of ordinary ones. **The two charts signalling in the same month is coherence evidence**, weak evidence against the two responses having moved for unrelated reasons, and it is offered as that rather than as a contribution.

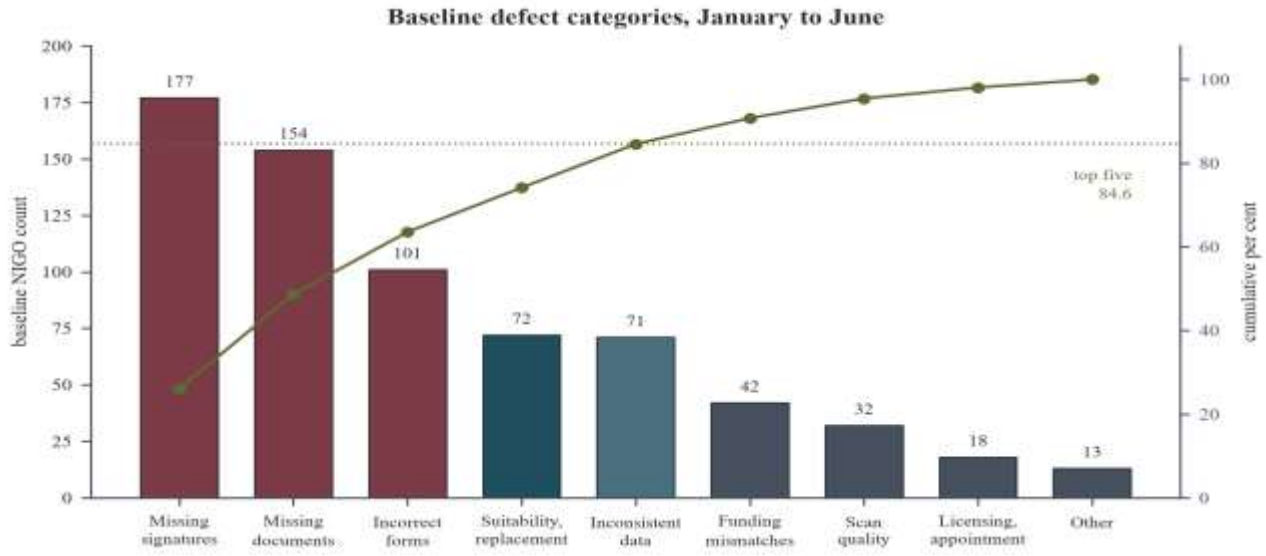


Figure 3. Baseline defect categories on the synthetic population. The three largest categories account for 63.5 per cent of baseline events and the five largest for approximately 84.6 per cent. The categories shaded differently are those addressed directly by capture redesign and point-of-entry validation; suitability and replacement, the fourth category, is the one Section VIII shows does not respond.

Overdispersion, checked rather than assumed. At census subgroups of this size a binomial p-chart is commonly overdispersed, which inflates false signals [10]. The baseline subgroup rates here are tightly clustered around the centre line, consistent with binomial variation rather than in excess of it, so the standard p-chart is retained. **Had the baseline shown excess variation, the p-prime chart would be the correct instrument**, and the protocol specifies that check before the chart is accepted rather than after a signal is disputed.

TABLE II. Comparison across the two periods, synthetic population

Measure	Baseline	Post-intervention	Change
Applications	1,825	1,875	
Not-in-good-order rate	37.3 per cent	18.1 per cent	**19.2 percentage points**
Mean cycle time	5.42 days	3.25 days	40.0 per cent reduction
Total rework	1,197.9 hours	500.3 hours	697.6 fewer hours
Mean rework per application	39.4 minutes	16.0 minutes	59.4 per cent reduction

On the arithmetic. The percentage-point change computes to 19.2 from the two stated endpoints. The rework reduction per application, at 59.4 per cent, is larger than the headline cycle-time figure and follows directly from the two period means.

Statistical testing, and what it does and does not establish. On the synthetic population the difference in proportions is significant on a two-proportion test, with a test statistic of approximately 13.1 computed from a pooled proportion of 0.275. Cycle-time and rework differences are significant on Welch's unequal-variance test [13], which is the correct form because the periods' variances differ by design. **These results demonstrate that the tests detect the shift present in the data at these subgroup sizes. They are a property of the design and the tests. They are not evidence about insurance operations**, and reporting a significant result on a shift that was injected by the generating mechanism would be circular if presented any other way.

VIII. The cause that did not move

This is the part of the protocol that the improvement literature does not cover, and it is the paper's primary contribution.

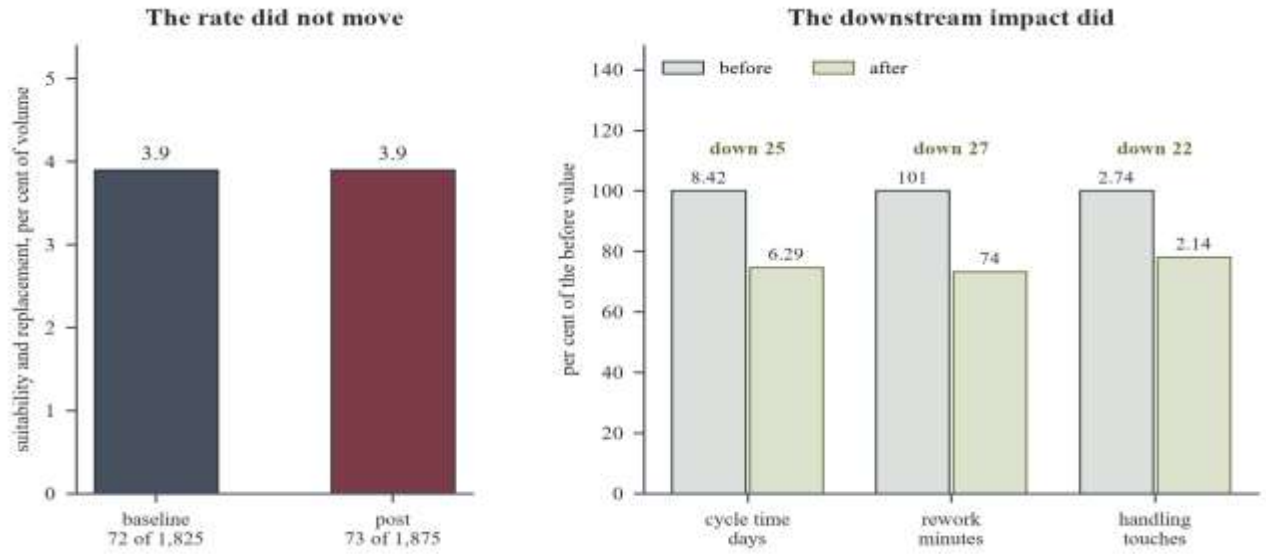


Figure 4. Suitability and replacement exceptions on the synthetic population. The incidence is unchanged across the two periods at approximately 3.9 per cent of volume, by construction. The downstream cost of each such exception is substantially reduced. The left panel is a parameter of the generating mechanism; the right panel is what the protocol does about it.

Why this category cannot be eliminated. Suitability and replacement exceptions depend on regulatory requirements, product-specific rules, customer information, or approvals outside the immediate operating team's control. No amount of capture redesign or point-of-entry validation removes a requirement imposed from outside the process.

The incidence, and its status. 72 of 1,825 in the baseline and 73 of 1,875 afterwards, approximately 3.9 per cent in both periods. **This flatness is a specified input to the generating mechanism, not an observation**, and it exists to exercise the protocol against a cause that does not respond. At these counts the plain normal approximation is marginal, and an adjusted interval for the difference of proportions [14] is the appropriate form if the comparison is to be reported at all.

TABLE III. Downstream cost of suitability and replacement exceptions, synthetic population

Measure	Before	After	Reduction
Mean cycle time for these cases	8.42 days	6.29 days	25 per cent
Mean rework per case	101 minutes	74 minutes	27 per cent
Mean handling touches	2.74	2.14	22 per cent

What was done instead of elimination. Available validation was moved upstream even where the exception itself could not be prevented; a dedicated exception path was created so these cases stopped competing with standard work; decision rules were clarified; cases were routed to designated specialists rather than to whoever was next; and escalation and aging thresholds were established so a case could not sit unattended.

The general form of the argument, which is the transferable part. An improvement programme framed around elimination has no move available when a cause is outside its control, and the usual outcome is that the category is quietly excluded from scope and the headline rate improves without the applicant's experience changing. **Partitioning the defect population into reducible and irreducible causes, and then targeting delay, rework and handoffs for the irreducible fraction, is a different and defensible response.** The improvement literature is framed almost uniformly around elimination and does not treat it; the nearest adjacent work is on workarounds and exception handling rather than anything in Lean Six Sigma or statistical process control.

What the numbers in Table III are. Parameters of the generating mechanism, chosen to represent a plausible effect of the exception-path changes. They illustrate the argument. They do not evidence it.

IX. Holding the gain

Ownership, stated precisely, because diffuse ownership is how control plans fail. The process owner owns the metric. Process Excellence maintains the chart and facilitates the response. **Operations owns the corrective action.** Separating the three prevents the common failure in which the function that maintains the chart is also expected to fix what the chart shows, and therefore does neither.

The control metric and its supports. The monthly not-in-good-order rate is primary, supported by first-pass yield, cycle time, rework minutes and the leading defect categories.

Three triggers, each catching a different failure shape. A monthly rate above the upper control limit requires immediate investigation. A sustained run of points above the centre line, even within the limits, triggers review for gradual deterioration, which is the shape a control chart's limits alone will not catch. And a material increase in a specific defect category or submission channel triggers targeted root-cause analysis **before** the overall rate deteriorates, which is the only one of the three that is preventive rather than reactive.

The reaction plan. On a trigger, the process owner stratifies by defect type, application type, submission channel and complexity; reviews recent process, form, staffing or system changes; and assigns corrective action with a documented owner and due date. Monthly Pareto reviews, quarterly defect-classification audits, standardised work instructions and change control over forms and validation rules run continuously.

The quarterly classification audit deserves emphasis because it is the one element that protects the measurement system over time. The definition repaired in Section IV will drift as staff turn over and as new product types appear, and a periodic re-audit of classification agreement is what detects that drift before it corrupts the metric the control plan depends on.

Why the control phase is framed as early warning rather than as maintenance. The author's stated aim is worth preserving directly: the goal is not simply to keep the rate low, but to create an early warning system so that deterioration is detected quickly, located precisely, and acted on before performance returns to its pre-improvement level. The systematic evidence on why continuous-improvement initiatives fail to hold [16] supports that framing.

X. Limitations

The data is synthetic and no result here is an operational finding. This is stated in the abstract, before Section I, in Section III, in every results caption and here. **No claim is made about any organisation, population or effect size.**

The observed reduction is the injected parameter recovered. Section VII reports it as such. A significant test on an injected shift demonstrates the test's sensitivity at the given design, nothing more.

The attribute measurement study is specified but not executed. This is the largest gap between protocol and demonstration. The generating mechanism does not model reviewer-level disagreement, so the measurement system repair that Section IV argues is the protocol's most important sequencing decision is supported by argument and by prior work in other domains [11], [12] rather than exercised here. **A natural next study** generates classifications from a modelled reviewer population with specified disagreement on borderline items, runs an agreement analysis, repairs the definition, and re-runs it, which would show whether the repair does what the protocol claims.

Twelve points is few for an individuals chart. Limits estimated from a six-point baseline are unstable, and Wheeler and Chambers are explicit about what a short baseline supports [7]. Generating more months would cost nothing and would be the obvious improvement to the design.

The variable control limits do little visible work at subgroup sizes ranging only from about 280 to 330, as Section VI states. They are correct rather than consequential.

The paired-chart design is textbook, not novel. Montgomery prescribes exactly this pairing [6]. It is reported as methodological rigour and as coherence evidence, not as a contribution.

Structured settlements are named in the title and are not grounded in literature. No refereed work on structured settlement operations was found. The domain framing rests on the author's practitioner knowledge and on the structure of the synthetic model and is presented that way.

No external validity. Nothing here transfers to any carrier, book of business or process without the study being redone on real data.

What a synthetic study can legitimately claim, and what this one claims: that the specified protocol detects a shift of known magnitude under stated conditions; that it is internally coherent and executable end to end, each stage's output feeding the next; and that its sensitivity to its own assumptions can be explored, which is the one respect in which a synthetic study is stronger than a case study, because a real deployment cannot be re-run.

XI. Conclusion

The protocol's first move is the one that matters, and it is not a process change.

Repairing the operational definition before measuring anything is what makes the rest interpretable, because a not-in-good-order rate is the output of a measuring instrument, and that instrument is a reviewer applying a definition. A baseline produced by an inconsistent classifier cannot anchor a comparison, and no process improvement will fix that. The published intake-defect work does not sequence it this way, and attribute agreement analysis, well established in manufacturing inspection, appears not to have been applied to insurance application completeness at all.

The second thing worth taking from this is the treatment of a cause that will not move. Suitability and replacement exceptions depend on requirements imposed from outside the process, and an improvement programme framed around elimination has nothing to offer them except quiet exclusion from scope. Partitioning the defect population and attacking delay, rework and handoffs for the irreducible fraction is the alternative, and it is the part of this protocol that the improvement literature does not cover.

What this paper is not is an account of an improvement that happened. Every figure in it was constructed for the study, the shift it detects was injected by the generating mechanism, and the statistical tests demonstrate their own sensitivity rather than anything about insurance. That form was chosen because the domain has no refereed literature at all and real operational data is almost never releasable, which leaves a reproducible method demonstration as the honest available contribution. The specification in Section III is complete enough to regenerate the dataset and re-run the analysis, which is the one thing a study on proprietary data cannot offer, and the measurement system study described in Section X is the work that would turn the protocol's central claim from an argument into a result.

Note on data, placed before anything else

Every quantity in this paper is synthetic. The author has no access to operational data from any organisation and did not use any. The dataset was constructed for this study. In the author's own words, recorded at intake: *"synthetic research data has been created and should not be represented as actual organizational results"*, and *"external disclosure approval is not applicable as this data is completely fabricated."*

What follows from that, applied throughout. Results sentences are anchored to the simulated process rather than to an organisation. The paper does not claim that a not-in-good-order rate was reduced anywhere. It claims that a specified protocol, exercised on a specified synthetic population, behaves in a stated way. No financial figure appears, because none would mean anything. The statistical tests reported in Section VII demonstrate that those tests detect the shift present in the data at the given subgroup sizes, which is a property of the tests and the design, not a fact about insurance operations.

On the title. The title names the study as simulated and its outcomes as modelled, which is what they are. An earlier formulation described the intervention and its outcomes without those qualifiers; it was changed because a title asserting measured outcomes over a methods section declaring constructed data would read as a case study carrying a disclaimer rather than as a method demonstration, and the two are not the same kind of paper.

Data availability. The data-generating mechanism is specified in Section III in enough detail to regenerate the dataset. For a synthetic study the generating specification is the data, and withholding it would defeat the purpose of declaring the study synthetic at all.

Attribution. Data validation by the author. External disclosure approval is not applicable, the data being wholly constructed for this study.

Confidentiality. No employer, carrier, customer or individual is named, and none is described in a way that would identify one.

References

[1] International Organization for Standardization, *Quantitative Methods in Process Improvement, Six Sigma, Part 1: DMAIC Methodology*, ISO 13053-1:2011, Geneva, Switzerland, Sep. 2011. [Online]. Available: https://www.iso.org/standard/52901.html (accessed Oct. 4, 2026).

[2] A. Sarkar, A. R. Mukhopadhyay, and S. K. Ghosh, "Improvement of claim processing cycle time through Lean Six Sigma methodology," *Int. J. Lean Six Sigma*, vol. 4, no. 2, pp. 171-183, 2013, https://doi.org/10.1108/20401461311319347.

[3] K. Sandner, S. Sieber, M. Tellermann, and F. Walthes, "A Lean Six Sigma framework for the insurance industry: Insights and lessons learned from a case study," *J. Bus. Econ.*, vol. 90, no. 5-6, pp. 845-878, 2020, https://doi.org/10.1007/s11573-020-00989-9.
[4] V. Sunder M., "Rejects reduction in a retail bank using Lean Six Sigma," *Prod. Planning Control*, vol. 27, no. 14, pp. 1131-1142, 2016, https://doi.org/10.1080/09537287.2016.1187312.
[5] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*. New York, NY, USA: D. Van Nostrand, 1931.
[6] D. C. Montgomery, *Introduction to Statistical Quality Control*, 8th ed. Hoboken, NJ, USA: Wiley, 2020.
[7] D. J. Wheeler and D. S. Chambers, *Understanding Statistical Process Control*, 3rd ed. Knoxville, TN, USA: SPC Press, 2010.
[8] International Organization for Standardization, *Control Charts, Part 2: Shewhart Control Charts*, ISO 7870-2:2023, 2nd ed., Geneva, Switzerland, Mar. 2023. [Online]. Available: https://www.iso.org/standard/78859.html (accessed Oct. 4, 2026).
[9] L. S. Nelson, "The Shewhart control chart, tests for special causes," *J. Qual. Technol.*, vol. 16, no. 4, pp. 237-239, 1984, https://doi.org/10.1080/00224065.1984.11978921.
[10] D. B. Laney, "Improved control charts for attributes," *Qual. Eng.*, vol. 14, no. 4, pp. 531-537, 2002, https://doi.org/10.1081/QEN-120003555.
[11] W. N. van Wieringen and J. de Mast, "Measurement system analysis for binary data," *Technometrics*, vol. 50, no. 4, pp. 468-478, 2008, https://doi.org/10.1198/004017008000000415.
[12] C. Marques, N. Lopes, G. Santos, I. Delgado, and P. Delgado, "Improving operator evaluation skills for defect classification using training strategy supported by attribute agreement analysis," *Measurement*, vol. 119, pp. 129-141, 2018, https://doi.org/10.1016/j.measurement.2018.01.034.
[13] B. L. Welch, "The generalization of Student's problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28-35, 1947, https://doi.org/10.1093/biomet/34.1-2.28.
[14] A. Agresti and B. Caffo, "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures," *Amer. Statistician*, vol. 54, no. 4, pp. 280-288, 2000, https://doi.org/10.1080/00031305.2000.10474560.
[15] T. A. Saurin, J. L. D. Ribeiro, and G. Vidor, "A framework for assessing poka-yoke devices," *J. Manuf. Syst.*, vol. 31, no. 3, pp. 358-366, 2012, https://doi.org/10.1016/j.jmsy.2012.04.001.
[16] R. S. McLean, J. Antony, and J. J. Dahlgaard, "Failure of continuous improvement initiatives in manufacturing environments: A systematic review of the evidence," *Total Qual. Manage. Bus. Excellence*, vol. 28, no. 3-4, pp. 219-237, 2017, https://doi.org/10.1080/14783363.2015.1063414.
[17] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statist. Med.*, vol. 38, no. 11, pp. 2074-2102, 2019, https://doi.org/10.1002/sim.8086.
[18] T. Monks, C. S. M. Currie, B. S. Onggo, S. Robinson, M. Kunc, and S. J. E. Taylor, "Strengthening the reporting of empirical simulation studies: Introducing the STRESS guidelines," *J. Simul.*, vol. 13, no. 1, pp. 55-67, 2019, https://doi.org/10.1080/17477778.2018.1442155.
[19] B. S. Siepe, F. Bartoš, T. P. Morris, A.-L. Boulesteix, D. W. Heck, and S. Pawel, "Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting," *Psychol. Methods*, advance online publication, 2024, https://doi.org/10.1037/met0000695.
[20] T. E. Raghunathan, "Synthetic data," *Annu. Rev. Statist. Appl.*, vol. 8, no. 1, pp. 129-140, 2021, https://doi.org/10.1146/annurev-statistics-040720-031848.