AMLR Asociația de Medicină de Laborator din România

Romanian Journal of Laboratory Medicine 2026;34(4):297–306

ORIGINAL RESEARCH

DOI:10.2478/rrlm-2026-0032

Three consecutive failures, one uncertain verdict: A simulation-based reassessment of accreditation-restriction rules in clinical chemistry

Radu Ilinca[1], Dan Adrian Luțescu[1]*, Ionela Ganea[2], Daniela Cristina Dimitriu[3]

1. Medical Informatics and Biostatistics, Faculty of Dentistry, "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania
2. 3rd Modern Languages, Faculty of Medicine, "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania
3. Biochemistry Department, Faculty of Medicine, "Grigore T. Popa" University of Medicine and Pharmacy, Iasi, Romania

ABSTRACT

Background: Proficiency testing (PT) is central to laboratory quality assurance; however, repeated unsatisfactory results may reflect both analytical performance and PT-scheme architecture. We evaluated whether three consecutive unsatisfactory results reliably identify deficient analytical performance in serum creatinine testing.

Methods: Monte Carlo simulation modelled three-round trajectories in an ISO 13528-based PT scheme. A total of 206 prespecified scenarios generated 1,446 analytical conditions varying laboratory performance, peer-group definition and size, heterogeneity, consensus construction, between-round persistence and σPT derivation. More than 875 million three-round trajectories were simulated. False-trigger and detection probabilities were evaluated for competent profiles and deficiencies caused by systematic bias, increased analytical imprecision or both.

Results: Under the reference condition, false-trigger probabilities were 0-0.03% for competent profiles, reaching 2.81% under the most adverse configuration examined. Detection probabilities were 20.47% for systematic bias, 0.38% for increased analytical imprecision and 16.24% for combined deficiency. With analytical performance held constant, changes in peer-group definition, focal-laboratory contribution to consensus or σPT derivation altered trigger probabilities by approximately 1.3- to 1.9-fold. Full between-round persistence increased false-trigger probability by approximately 24% relative to independence. Distributional sensitivity analysis produced a maximum absolute difference of 0.69 percentage points.

Conclusions: Three consecutive unsatisfactory PT results are not an invariant indicator of deficient analytical performance. Their probability depends on both the mechanism of deficiency and PT-scheme architecture. Such outcomes should therefore trigger investigation rather than independently establish analytical deficiency or its cause.

Keywords: analytical imprecision, external quality assessment, laboratory accreditation, proficiency testing, serum creatinine

Received: 3 September 2026; Accepted: 2 October 2026; Published: 29 October 2026.

INTRODUCTION

Proficiency testing (PT), commonly implemented in medical laboratories through external quality assessment (EQA), is a central component of laboratory quality assurance and accreditation under ISO 15189 [1] and ISO/IEC 17043 [2]. International surveys confirm its widespread use and document persistent unacceptable EQA performance across laboratory systems [3-5]. This is particularly relevant for serum creatinine, for which method- and platform-dependent differences remain documented [6, 7] and may affect eGFR estimates and CKD-category assignment [8-10].

However, a PT classification reflects both laboratory performance and the statistical framework used to evaluate it. Peer-group definition and size, consensus estimation and the standard deviation for proficiency assessment can influence the resulting classification [11-15]. Peer-group partitioning can materially alter estimated agreement and bias between analytical systems [12], a relevant consideration for creatinine given persistent systematic differences among Jaffe-based platforms [6, 7]. Therefore, a satisfactory or unsatisfactory PT result cannot be considered entirely independent of the architecture of the scheme producing it.

This becomes consequential when repeated PT outcomes inform accreditation decisions. ISO 15189 requires investigation of EQA results outside specified criteria and assessment of the significance of the resulting nonconformity [1] while

* Corresponding author: Dan Adrian Luțescu, Medical Informatics and Biostatistics, Faculty of Dentistry, "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania. Email: dan-adrian.lutescu@drd.umfcd.ro

ISO/IEC 17043 recognises limitations in using PT results alone to determine competence [2,16]. Correspondingly, ISO/IEC 17011 requires accreditation decision-makers to understand risk-based assessment principles and to be competent to evaluate assessment outcomes [17]. Nevertheless, a fixed rule based on three consecutive unsatisfactory results is administratively attractive because it is simple and auditable. Its apparent objectivity does not establish that three identical classifications necessarily constitute three equivalent pieces of evidence about the laboratory itself. If the probability of such a sequence changes while laboratory performance remains unchanged, the resulting decision cannot be attributed to laboratory competence alone.

The operating characteristics of repeated-outcome PT rules remain insufficiently characterised. We therefore used Monte Carlo simulation to evaluate a three-consecutive-unsatisfactory-result ("three-strikes") rule across predefined competent and deficient serum-creatinine profiles. We quantified false-trigger probabilities, detection of deficiencies caused by systematic bias, increased analytical imprecision or both, and the effects of peer-group definition and size, consensus construction, between-round persistence and derivation of the standard deviation for proficiency assessment. Therefore, the study tested whether three consecutive unsatisfactory results can function as a self-sufficient indicator of analytical deficiency or should instead be interpreted as a trigger for further investigation.

METHODS

Ethics statement

Ethical approval was not required for this study.

Study design

This study used Monte Carlo computer simulation to evaluate the operating characteristics of a fixed "three consecutive unsatisfactory results" accreditation-restriction rule ("three-strikes rule") applied to serum creatinine measured within a proficiency-testing (PT) scheme constructed according to ISO 13528 [18]. The simulation evaluated whether three consecutive unsatisfactory PT results (FFF) reliably distinguished predefined competent from analytically deficient laboratory-performance profiles and how this probability was affected by characteristics of the PT scheme itself.

The complete simulation architecture, from analytical-performance profile and peer-group generation through focal-laboratory result generation, Algorithm A consensus estimation, z-score classification, three-round persistence and probability estimation, is summarised in Figure 1.

Analytical-performance profiles

The analytical performance of the evaluated ("focal") laboratory was parameterised using declared performance data from six representative suppliers of Jaffe-method creatinine analytical systems on the Romanian market (combined market share ≈80%), together with a published biological-variation-based analytical-performance specification for creatinine [19,20].

Eight ground-truth analytical-performance profiles were defined a priori along two independently parameterised axes: systematic bias and analytical imprecision, the latter expressed as coefficient of variation (CV). Full numerical parameterisation and the corresponding ground-truth classifications are provided in Table 1A. The competent profiles represented zero-bias, near-limit positive- and negative-bias and optimal-imprecision conditions. A grey-zone profile was positioned at the minimum imprecision specification and excluded from the primary competent-versus-deficient analysis. Deficient profiles separately represented excessive systematic bias, excessive analytical imprecision and their combination, allowing the two analytical failure mechanisms to be evaluated independently and jointly.

The optimal, desirable and minimum analytical CV specifications were 1.075%, 2.150% and 3.225%, respectively; the corresponding bias specifications were 1.699%, 3.397% and 5.096%. For the deliberately deficient profiles, the relevant analytical component was set at 150% of its minimum specification: 7.644% for systematic bias and 4.838% for analytical imprecision. Realistic bounds for method- and system-level offsets were additionally anchored to declared precision and calibrator-traceability data from six representative creatinine reagent/calibrator manufacturers active on the Romanian market, as well as to published multicentre creatinine comparison data documenting between-vendor deviations of up to approximately 10% from reference-method-assigned values [20].

Scenario design

A scenario was defined as a fully specified combination of analytical-performance profile and PT-scheme characteristics. The principal experimental factors included peer-group definition and size, method/system and material/matrix effects, peer-group heterogeneity, between-round persistence, focal-laboratory contribution to consensus and σPT derivation. Their levels and analytical purposes are summarised in Table 1B.

A total of 206 base scenarios were defined by the prespecified experimental design. Orthogonal secondary design choices expanded these to 1,446 distinct analytical conditions submitted to Monte Carlo evaluation. Peer groups represented all-participant, method-level or analytical-system-level classifications, with group sizes ranging from 10 to 1,000 participants depending on scenario. Method-level offsets were evaluated bidirectionally from -10% to +10%, while analytical-system and material/matrix effects ranged from -4% to +4%.

As a contextual, non-normative anchor for realistic peer-group sizes and result dispersion, a single anonymised proficiency-testing report for creatinine from a Romanian accredited PT provider was examined, reporting both a pooled all-method evaluation and a narrower method-specific evaluation [12]. This report was not treated as ground truth and was not used to define competence status or to determine any structural parameter of the simulation; it served exclusively to verify that the simulated peer-group configurations correspond to conditions observed in practice.

Round-level simulation, consensus estimation and classification

For each simulated PT round, peer-laboratory results were first generated according to the specified peer-group definition, size, heterogeneity, method/system offsets and material/matrix effects. The focal-laboratory result was generated separately, according to its predefined ground-truth analytical-performance profile, with systematic bias and analytical imprecision applied independently; where specified by the scenario, a persistent component was carried across consecutive rounds.

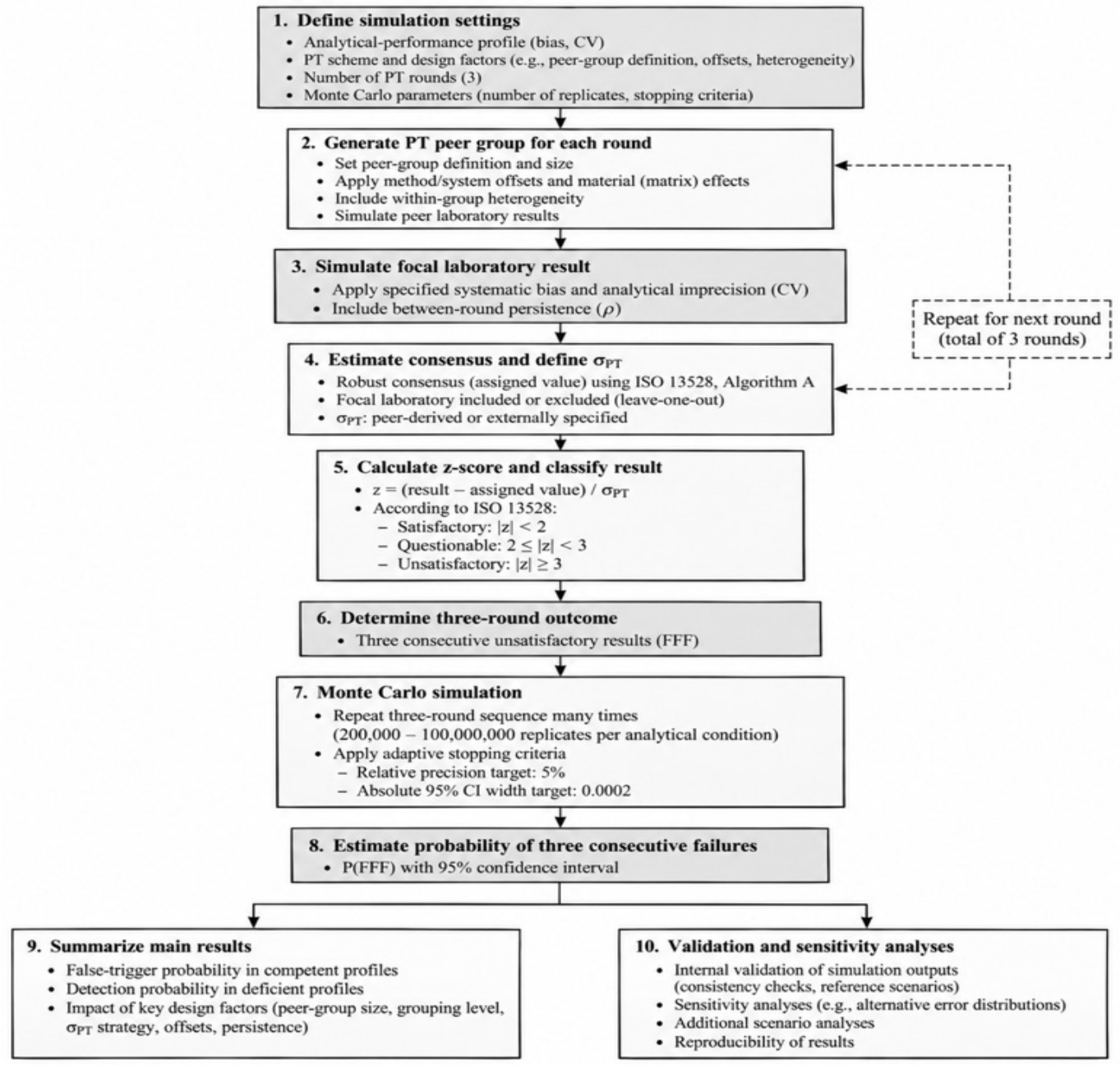


Figure 1. Methodological workflow of the Monte Carlo simulation

Each analytical condition proceeds from the definition of simulation settings (analytical-performance profile, PT-scheme design factors, number of rounds and Monte Carlo parameters) through generation of the simulated peer group and focal-laboratory result, robust consensus estimation using ISO 13528 Algorithm A, z-score classification and determination of the three-round outcome. The round-level procedure is repeated across three consecutive PT rounds while incorporating the specified degree of between-round persistence. Complete three-round sequences are then replicated adaptively to estimate the probability of three consecutive unsatisfactory results, P(FFF), with its 95% confidence interval. Subsequent analyses summarise false-trigger and detection probabilities and evaluate the influence of PT-design factors, together with internal validation and sensitivity analyses.

Table 1A. Analytical-performance profiles and principal design factors used in the Monte Carlo simulation: Ground-truth analytical-performance profiles

Analytical performance profile	Systematic bias (%)	Analytical imprecision CV (%)	Ground-truth classification	Parameterization
Competent, zero bias	0.000	2.150	Competent	Zero bias; CV at the desirable performance specification
Competent, near-positive bias limit	+3.397	2.150	Competent	Bias and CV at the desirable performance specifications
Competent, near-negative bias limit	-3.397	2.150	Competent	Negative counterpart of the near-positive-limit profile
Competent, optimal imprecision	0.000	1.075	Competent	Zero bias; CV at the optimal performance specification
Grey-zone imprecision	0.000	3.225	Grey zone (excluded from the primary competent-vs-deficient analysis)	Zero bias; CV at the minimum performance specification
Deficient by bias	+7.644	2.150	Deficient	Bias at 150% of the minimum bias specification; CV at the desirable specification
Deficient by imprecision	0.000	4.838	Deficient	Zero bias; CV at 150% of the minimum imprecision specification
Deficient by both	+7.644	4.838	Deficient	Bias and CV each at 150% of their respective minimum specifications

Table 1B. Analytical-performance profiles and principal design factors used in the Monte Carlo simulation: Principal PT-scheme and the Monte Carlo design factors

Design factor	Levels of implementation	Analytical purpose
Peer-group definition	All participants; method-level;analytical-system-level	Assess the effect of increasing peer-group specificity
Peer-group size (N)	All participants: 200, 500, 1000; method-level: 50, 100, 250, 500; system-level: 10, 15, 25, 40, 75	Assess the effect of peer-group size
Method-level offset	-10,-7.5,-5,-2.5,-1, 0, +1, +2.5, +5, +7.5, +10%	Assess bidirectional between-method effects
System-level offset	-4,-2,-1, 0, +1, +2, +4%	Assess analytical-system-specific effects
Material/matrix effect	-4,-2,-1, 0, +1, +2, +4%	Assess persistent material-related effects
Between-round persistence (ρ)	0, 0.25, 0.50, 0.75, 1.00	Assess conditions ranging from independence to complete persistence across rounds
Peer-group heterogeneity	Homogeneous; moderate; strong	Assess increasing within-peer-group heterogeneity
Focal laboratory contribution to consensus	Included; excluded (leave-one-out)	Assess the effect of focal-laboratory self-influence on the consensus
SD for proficiency assessment (σPT)	Peer-derived; externally specified	Assess sensitivity to the method used to define σPT
Consensus estimation	ISO 13528 Algorithm A	Obtain the robust assigned value and associated dispersion
Number of consecutive PT rounds	3	Represent the three-consecutive-failure (FFF) rule
Primary simulation endpoint	P(FFF)	Estimate false-trigger probability for competent profiles and detection probability for deficient profiles
Monte Carlo sampling	Adaptive; 200,000 to 100,000,000 three-round sequences per analytical condition	Achieve prespecified numerical precision while limiting unnecessary computation
Monte Carlo precision criteria	Relative precision target: 5%; absolute confidence-interval width target: 0.0002	Prespecified adaptive stopping criteria

Analytical-performance specifications were defined independently of the PT-provider data, using published biological-variation-based criteria [19,20]. For creatinine, the underlying within-subject and between-subject biological variation were 4.3% and 12.89%, respectively; the resulting optimal, desirable and minimum analytical CV specifications were 1.075%, 2.150% and 3.225%, while the corresponding bias specifications were 1.699%, 3.397% and 5.096%. Deficient profiles were deliberately positioned beyond the minimum specification by setting the relevant analytical component(s) to 150% of the corresponding minimum limit. The grey-zone profile was positioned at the minimum imprecision specification with zero bias and was excluded from the primary competent-versus-deficient classification, but retained in a secondary tiered-gradient analysis. The analytical imprecision was parameterized as CV and was treated separately from calibrator uncertainty, PT-material uncertainty and the uncertainty of the PT assigned value, u(xPT). **Abbreviations:** CV – coefficient of variation, FFF – three consecutive unsatisfactory PT results, PT – proficiency testing, σPT – standard deviation for proficiency assessment.

After generation of the peer-group and focal-laboratory results, robust consensus estimation reproduced ISO 13528 Algorithm A [18]. Depending on the experimental condition, the focal laboratory was either included in the consensus population or excluded using a leave-one-out approach, allowing the influence of the focal laboratory's own result on the consensus against which it was subsequently evaluated to be quantified.

Algorithm A provided the robust assigned value, x*, and robust standard deviation, s*. The standard uncertainty of the assigned value was calculated by using the ISO 13528 expression:

$$u(x^*) = (1.25 \times s^*)/\sqrt{n} \quad (1)$$

where n is the number of participants in the comparison group.

The focal laboratory's simulated result (xlab) was then compared against this group consensus by using the standard z-score:

$$z = (x_lab - x^*)/\sigma_PT \quad (2)$$

where σPT denotes the standard deviation for proficiency assessment. Values with $|z| < 2$ were classified as satisfactory, $2 \leq |z| < 3$ as questionable and $|z| \geq 3$ as unsatisfactory.

Two approaches to σPT were evaluated. In the primary approach, σPT was derived from the simulated peer-group consensus. In a separate comparison, σPT was externally specified using a target value independent of the simulated peer-group dispersion, allowing the extent to which FFF probabilities depend on the definition of σPT itself to be quantified [13].

Treatment of bias, imprecision and uncertainty

The systematic bias and the analytical imprecision were treated as distinct, independently parameterised quantities throughout the simulation. The analytical imprecision was parameterised as CV and represented the random dispersion of laboratory results around the corresponding analytical expectation. Neither bias nor imprecision was used interchangeably with measurement uncertainty.

The uncertainty components represented in the simulation were structurally separated from these two analytical-performance dimensions. The calibrator uncertainty, the PT-material uncertainty and the uncertainty of the PT assigned value, u(x*), were represented as conceptually distinct quantities and were not substituted for either systematic bias or analytical imprecision, consistent with the terminology and the conceptual framework of JCGM [21]. Calibrator uncertainty was informed by the uncertainty values reported in the manufacturers' calibration traceability certificates.

Longitudinal structure and persistence

Each analytical condition comprised three consecutive PT rounds, the minimum longitudinal sequence required to evaluate the three-consecutive-unsatisfactory-result rule. Because the analytical influences may persist across successive PT rounds rather than arise independently at each occasion, the between-round persistence was explicitly incorporated into the model. The persistence parameter (ρ) was evaluated at 0, 0.25, 0.50, 0.75 and 1.00, representing conditions ranging from round-to-round independence to complete persistence. The effect of persistence on the probability of FFF was quantified using the amplification factor:

$$A = P(F1 \cap F2 \cap F3) / [P(F1) \times P(F2) \times P(F3)] \quad (3)$$

where F1, F2 and F3 denote unsatisfactory outcomes in the first, second and third PT rounds, respectively. An amplification factor greater than 1 indicates that the observed probability of three consecutive failures exceeds that expected from the product of the marginal round-level failure probabilities under an independence assumption.

As an internal implementation check, an artificially introduced method-level offset produced the expected monotonic, approximately symmetric response around zero deviation across all tested conditions- used to identify potential directional or implementation artefacts, not as independent evidence of clinical or metrological causality.

Monte Carlo procedure

For each of the 1,446 analytical conditions, complete three-round sequences were generated repeatedly by the Monte Carlo simulation. The number of replicates was determined adaptively rather than uniformly, allocating greater computational effort to rare outcomes. The simulation proceeded in successive batches, with a minimum of 200,000 and a maximum of 100,000,000 complete three-round sequences per condition. The adaptive stopping used prespecified numerical-precision criteria - a relative precision target of 5% and an absolute 95% confidence-interval width criterion of 0.0002 - allowing stable estimation of low event probabilities without requiring the maximum replicate count where adequate precision was reached earlier.

The primary endpoint was the probability of three consecutive unsatisfactory results, P(FFF). For ground-truth competent profiles, P(FFF) was interpreted as the false-trigger probability of the three-strikes rule; for ground-truth deficient profiles, the same quantity represented the probability that the rule detected the predefined analytical deficiency. Confidence intervals were calculated for all estimated probabilities.

The computational workflow was organised into sequential stages - scenario generation, the Monte Carlo simulation, stratified analysis and production of summary tables and figures - each consuming only the finalised output of the preceding stage, with no manual adjustment of simulated results or calculated probabilities between scenario definition and final analysis [22].

Sensitivity and internal validation

The sensitivity to the assumed distribution of analytical measurement noise was examined by using a reduced, prespecified grid of representative conditions, comparing the primary Normal-distribution assumption against a right-skewed lognormal distribution matched for mean and variance, to evaluate whether the principal findings depended materially on the assumed shape of the random-error distribution.

The implementation of ISO 13528 Algorithm A and the calculation of assigned-value uncertainty, u(x*), were additionally validated against independently sourced, real-world creatinine PT data from a Romanian accredited provider, comprising both all-method and Jaffe-method peer-group summaries; this validation reproduced the

reported assigned-value uncertainty within the expected rounding tolerance. This dataset was used exclusively to verify numerical implementation and the plausibility of simulated peer-group characteristics, never to assign competence status to simulated laboratories or to determine any structural model parameter.

RESULTS

Across the 1,446 analysed conditions, the probability of three consecutive unsatisfactory results varied according to both the ground-truth analytical-performance profile and the configuration of the PT scheme. The results are first presented for the reference condition and subsequently for the principal PT-scheme design factors.

Reference condition

The reference condition comprised a system-level peer group of 25 participants, peer-derived σPT and inclusion of the focal laboratory in the round consensus. Table 2 presents the probability of three consecutive unsatisfactory results, P(FFF), for each primary analytical-performance profile.

No FFF event was observed for the zero-bias or optimal-imprecision competent profiles during 400,000 simulated three-round cycles. The competent profiles positioned near the positive and negative acceptability limits each had P(FFF) = 0.03%. Across all simulated conditions, the highest false-trigger probability observed among competent profiles was 2.81%, corresponding to approximately one event per 36 three-round cycles.

For deficient profiles under the reference condition, P(FFF) was 20.47% for systematic bias alone, 0.38% for elevated imprecision alone and 16.24% for combined bias and imprecision (Figure 2). The trigger probability for the bias-only profile was therefore approximately 54 times that of the imprecision-only profile. Across all conditions examined for the imprecision-only profile, detection did not exceed approximately 1.4%.

Effects of PT-scheme design

The peer-group definition, focal-laboratory contribution to consensus and σPT derivation substantially changed P(FFF) despite unchanged ground-truth analytical performance (Table 3).

By changing from a system-level peer group (N=25) to an all-participants group (N=500) increased P(FFF) from 20.47% to 35.63% for the bias-only profile, from 0.38% to

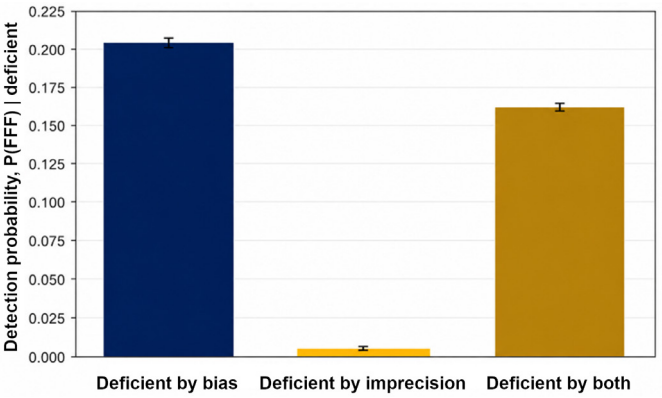


Figure 2. Detection probability by deficient-profile mechanism (system grouping)

The probability of three consecutive unsatisfactory results for deficient analytical-performance profiles under the reference condition, shown separately for systematic-bias-only, imprecision-only and combined bias-and-imprecision deficiency. Error bars represent 95% Wilson confidence intervals.

Table 2. The three-round trigger probability by analytical-performance profile under the reference condition (system-level peer group, N = 25, peer-derived σPT, focal laboratory included in the consensus)

Profile	Ground truth	P(FFF)	Approximate frequency
Zero systematic bias	Competent	0.00%	Not observed in 400,000 cycles
Near-positive acceptability limit	Competent	0.03%	~1 in 3,300
Near-negative acceptability limit	Competent	0.03%	~1 in 3,300
Optimal imprecision	Competent	0.00%	Not observed in 400,000 cycles
Systematic bias only	Deficient	20.47%	~1 in 5
Elevated imprecision only	Deficient	0.38%	~1 in 263
Combined bias + imprecision	Deficient	16.24%	~1 in 6

Abbreviations: P(FFF) – probability of three consecutive unsatisfactory PT results, PT – proficiency testing, σPT – standard deviation for proficiency assessment.

Table 3. The effect of selected PT-scheme design characteristics on the probability of FFF

Design comparison	Deficiency mechanism	Condition A	Condition B	P(FFF), A	P(FFF), B	B/A
Peer-group definition	Systematic bias only	System (N=25)	All participants (N=500)	20.47%	35.63%	1.74×
Peer-group definition	Elevated imprecision only	System (N=25)	All participants (N=500)	0.38%	0.61%	1.61×
Peer-group definition	Combined	System (N=25)	All participants (N=500)	16.24%	21.49%	1.32×
Focal-laboratory consensus contribution	Systematic bias only	Included	Excluded	20.47%	31.51%	1.54×
Focal-laboratory consensus contribution	Elevated imprecision only	Included	Excluded	0.38%	0.72%	1.89×
σPT derivation	Systematic bias only	Peer-derived	Externally specified	20.47%	31.65%	1.55×
σPT derivation	Elevated imprecision only	Peer-derived	Externally specified	0.38%	0.52%	1.37×

Abbreviations: P(FFF) – probability of three consecutive unsatisfactory PT results, PT – proficiency testing, σPT – standard deviation for proficiency assessment.

0.61% for the imprecision-only profile and from 16.24% to 21.49% for the combined profile. By excluding the focal laboratory from consensus estimation increased the bias-only probability from 20.47% to 31.51% and the imprecision-only probability from 0.38% to 0.72%. By using externally specified rather than peer-derived σPT increased the corresponding probabilities to 31.65% and 0.52%, respectively.

The peer-group size interacted with consensus construction. For the combined-deficiency profile within a system-level peer group, increasing N from 10 to 75 increased amplified P(FFF) from 9.3% to 19.7% when the focal laboratory contributed to the consensus, whereas the corresponding change was only approximately 20% to 21% when it was excluded.

The complete between-round persistence produced an amplification factor of approximately 1.24 for the competent profile relative to the probability expected under independence. For deficient profiles, persistence produced little additional amplification.

The distributional sensitivity analysis yielded similar estimates under Normal and mean-and-variance-matched right-skewed lognormal noise models. Across the representative conditions examined, the mean relative difference in P(FFF) was 11.5% and the maximum absolute difference was 0.69 percentage points.

Overall range of outcomes

The ranges observed across the simulated conditions are summarised in Table 4. Across all conditions, competent profiles had false-trigger probabilities between 0 and 2.81%. Non-detection ranged from 56.2% to 88.4% for systematic-bias-only deficiency and from 98.6% to 99.9% for imprecision-only deficiency. For combined deficiency, non-detection ranged from 28.0% to 100%.

Overall, P(FFF) depended on both the underlying analytical-performance profile and the structural characteristics of the PT scheme.

Table 4. Observed ranges of three-round false-trigger and non-detection probabilities across simulated conditions

Laboratory status	Metric	Range
Competent (any profile)	False-trigger probability	0.00%–2.81%
Deficient-systematic bias only	Non-detection probability	56.2%–88.4%
Deficient-elevated imprecision only	Non-detection probability	98.6%–99.9%
Deficient-combined mechanism	Non-detection probability	28.0%–100%

DISCUSSION

The principal finding of this study concerns not only how frequently the three-strikes rule triggers, but what such a trigger can reasonably be taken to mean. The Guide to the Expression of Uncertainty in Measurement distinguishes measurement error from measurement uncertainty as different concepts [21], a distinction whose application in laboratory medicine has itself been discussed [23]. A PT z-score describes the position of a laboratory result relative to an assigned value and a specified standard deviation for proficiency assessment; it is not, by itself, a direct measurement of the underlying analytical error within that laboratory. The present simulations show why this distinction matters: the probability of three consecutive unsatisfactory results depended both on the mechanism of analytical deficiency and on the construction of the peer comparison, consensus and σPT.

For competent analytical-performance profiles, false triggers were uncommon under the reference condition. However, across the wider design space, the maximum false-trigger probability reached 2.81% under the most adverse configuration examined. This value represents the upper bound of the simulated conditions rather than an expected rate for routine PT practice. Nevertheless, it demonstrates that three consecutive unsatisfactory outcomes can arise in a laboratory whose predefined analytical-performance profile remains within the specified limits. This observation is compatible with the broader empirical finding that unacceptable EQA outcomes vary across schemes and settings even among laboratories operating within established quality frameworks [5].

The converse result was more pronounced. Under the reference condition, systematic-bias-only deficiency produced an FFF probability of 20.47%, compared with 0.38% for imprecision-only deficiency, an approximately 54-fold difference. Across all conditions examined, detection of the imprecision-only profile did not exceed approximately 1.4%. Within the simulated conditions, the three-strikes rule was therefore substantially more sensitive to systematic bias than to increased analytical imprecision and did not provide a mechanism-independent assessment of analytical performance. This asymmetry is consistent with creatinine method-comparison studies documenting substantial vendor- and method-dependent systematic differences among Jaffe-based analytical systems [6, 7].

The interpretation of this finding should remain within the scope of the simulation. No patient-level creatinine values, eGFR estimates, CKD-stage transitions or clinical outcomes were modelled in this study. Consequently, the present results demonstrate differences in the statistical operating characteristics of the PT rule, not clinical harm arising from undetected analytical deficiency. The potential clinical relevance of creatinine measurement error is supported independently by the eGFR and CKD-classification literature [9, 10], but translating the probabilities reported here into patient-level consequences would require a separate analysis.

An additional finding was that PT-scheme construction changed the probability of FFF while the focal laboratory's underlying analytical performance was held constant. Under the reference comparisons, changing peer-group definition, excluding the focal laboratory from consensus estimation or replacing peer-derived σPT with an externally specified value changed trigger probabilities by approximately 1.3- to 1.9-fold, depending on the analytical profile. Empirical EQA studies have similarly shown that peer-group partitioning can alter the estimated relationship between an individual analytical procedure and its comparison group [12]. More generally, this supports previous calls for laboratories and EQA providers to understand the statistical and organisational characteristics that influence EQA interpretation [24].

The peer-group size showed a related interaction with consensus construction. Increasing the number of participants had a marked effect when the focal laboratory contributed to the consensus, but a much smaller effect when it was excluded. This behaviour is consistent with the greater influence that an individual observation can exert on consensus estimation in smaller comparison groups [11]. It should not be interpreted as a failure of Algorithm A, which is designed to provide robust estimates in the presence of discrepant observations. Rather, it illustrates that the operating characteristics of a subsequent classification rule depend not only on the robust estimator itself, but also on how the comparison group and consensus procedure are configured.

The comparison between peer-derived and externally specified σPT reinforces this point. For an unchanged bias-only laboratory profile, the FFF probability increased from 20.47% to 31.65% when externally specified rather than peer-derived σPT was used. The corresponding increase for the imprecision-only profile was smaller, but remained detectable. Thus, the standard deviation used for proficiency assessment is not merely a computational detail when repeated categorical outcomes are subsequently used as evidence in an accreditation decision. Alternative approaches to consensus and uncertainty treatment have been proposed in the metrological literature [13], although their comparative performance was not evaluated in the present simulation.

The between-round persistence provided a further qualification to the interpretation of repeated failures. The complete persistence increased the FFF probability for the competent profile by approximately 24% relative to the independence benchmark, whereas little additional amplification was observed for already deficient profiles. Three consecutive PT outcomes should therefore not automatically be interpreted as three statistically independent confirmations of an underlying condition. At the same time, sensitivity analysis using a matched right-skewed lognormal distribution produced only limited absolute changes in the estimated probabilities, suggesting that the principal findings were not driven by the Normal-distribution assumption.

These findings also have implications for the information accompanying a PT classification. A satisfactory or unsatisfactory designation is clearly a simplified summary of a statistical comparison. The simulations show that its operating characteristics can depend on peer-group definition, consensus inclusion, σPT derivation, group size and persistence, none of which is conveyed by the categorical label itself. Greater transparency regarding the statistical construction underlying a PT result may therefore be particularly important when repeated classifications are subsequently used for decisions extending beyond the immediate PT round. Performance-based quality-control strategies similarly adapt statistical decision rules to the underlying analytical performance [25]. This interpretation is consistent with the broader emphasis in ISO 15189 and the EQA literature on appropriate use and interpretation of external quality-assessment information [1, 2, 24, 26].

Several limitations should be considered. First, the study is simulation-based and necessarily represents a defined set of analytical-performance profiles and PT-scheme configurations rather than the full diversity of laboratory practice. Second, serum creatinine measured by Jaffe-based systems was used as the analytical model; the numerical probabilities should therefore not be assumed to apply unchanged to other measurands, methods or PT schemes. Third, the ground-truth profiles were deliberately constructed to isolate systematic bias and analytical imprecision, whereas real laboratory deficiencies may contain multiple interacting components. Fourth, although a broad parameter space was examined, the simulation cannot reproduce every operational feature of an individual PT provider or accreditation system. Finally, no patient-level outcomes were simulated and clinical consequences cannot be inferred quantitatively from the present probabilities.

The results should therefore not be interpreted as an argument against proficiency testing, robust consensus estimation or repeated-outcome-based accreditation oversight. EQA remains an essential component of laboratory quality assurance [3, 4]. Rather, the findings distinguish the evidential roles of screening and attribution. Three consecutive unsatisfactory PT results provide a defensible signal that the underlying analytical process warrants investigation, but the simulations do not support treating that sequence, by itself, as proof of a particular analytical deficiency or its cause. In this sense, the appropriate interpretation of an FFF outcome is a trigger for investigation rather than completion of the investigation.

CONCLUSIONS

Across nearly 1,500 simulated analytical conditions, the probability of three consecutive unsatisfactory PT results depended not only on the laboratory's underlying analytical performance, but also on the mechanism of deficiency and on the statistical architecture of the PT scheme itself. False triggers remained rare for competent profiles, while imprecision-driven deficiency was detected far less reliably than bias-driven deficiency of comparable severity. Scheme-design choices altered the probability of an identical outcome by a margin comparable to, and at times exceeding, the deficiency itself, despite unchanged laboratory performance. A fixed criterion such as "three consecutive unsatisfactory rounds" therefore does not correspond to a single operating-risk profile but to a family of profiles, rarely visible to either the laboratory or the body applying the rule.

For accreditation bodies, a mechanical three-strikes criterion applied without reference to the underlying scheme architecture cannot by itself satisfy the risk-based evaluation already required by ISO/IEC 17043, ISO 15189, and ISO/IEC 17011: such a sequence should trigger a documented investigation integrating the PT findings with independent evidence of performance, not function as an automatic verdict. For PT/EQA providers, the same results argue for disclosing the peer-group definition, consensus construction and σPT-derivation method behind a given classification alongside the satisfactory/unsatisfactory label, since these choices can shift classification probability by a margin comparable to a genuine deficiency.

Three consecutive unsatisfactory results are best treated as the start of an investigation, not as its conclusion.

ABREVIATIONS

CKD – Chronic Kidney Disease

eGFR – estimated Glomerular Filtration Rate

EQA – External Quality Assessment

FFF – three consecutive PT unsatisfactory outcomes (Fail–Fail–Fail)

GUM – Guide to the Expression of Uncertainty in Measurement

IFCC – International Federation of Clinical Chemistry and Laboratory Medicine

ISO – International Organization for Standardization

JCGM – Joint Committee for Guides in Metrology

PT – Proficiency Testing

ACKNOWLEDGEMENT

None.

AUTHORS' CONTRIBUTION

RI: conceptualization, methodology, software, data curation.

DAL: software, investigation, formal analysis, visualization.

IG: writing (original draft), writing (review & editing)

DCD: supervision, project administration.

All authors have read and approved the final version of the manuscript.

CONFLICT OF INTEREST

None to declare.

Publisher's Note: Editorial Office stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

REFERENCES

1. International Organization for Standardization. ISO 15189:2022. Medical laboratories — Requirements for quality and competence. Standard. Geneva, Switzerland: International Organization for Standardization; 2022.
2. International Organization for Standardization. ISO/IEC 17043:2023. Conformity assessment — General requirements for the competence of proficiency testing providers. Standard. Geneva, Switzerland: International Organization for Standardization; 2023.
3. Wheeler SE, Blasutig IM, Dabla PK, Giannoli JM, Vassault A, Lin J, et al. Quality standards and internal quality control practices in medical laboratories: an IFCC global survey of member societies. Clin Chem Lab Med. 2023;61(12):2094-101. DOI: 10.1515/cclm-2023-0492
4. Blasutig IM, Wheeler SE, Bais R, Dabla PK, Lin J, Perret-Liaudet A, et al. External quality assessment practices in medical laboratories: an IFCC global survey of member societies. Clin Chem Lab Med. 2023;61(8):1404-10. DOI: 10.1515/cclm-2023-0057
5. Bais R, Vassault A, Blasutig IM, Dabla PK, Lin J, Perret-Liaudet A, et al. External quality assessment performance in ten countries: an IFCC global laboratory quality project. Clin Chem Lab Med. 2024;62(12):2435-43. DOI: 10.1515/cclm-2024-0859
6. Bargnoux A-S, Kuster N, Cavalier E, Piéroni L, Souweine J-S, Delanaye P, et al. Serum creatinine: advantages and pitfalls. Journal of Laboratory and Precision Medicine. 2018;3:71-. DOI: 10.21037/jlpm.2018.08.01
7. Carobene A, Vidali M, Ceriotti F, Alberghini A, Arcangeli A, Bellandi W, et al. Enzymatic creatinine versus alkaline picrate: evidence from a multicenter human-pool study. Journal of Laboratory and Precision Medicine. 2026;11:5-. DOI: 10.21037/jlpm-2025-1-59
8. Preda EC, Oprea OR, Barabas AZ, Fotache AM, Dobreanu M. When equations disagree: the impact of creatinine-based eGFR in CKD diagnosis and reclassification. BMC Nephrol. 2025;26(1):647. DOI: 10.1186/s12882-025-04584-4
9. Osredkar J, Klemencic I, Kumer K, Pajek J, Knap B. Comparison of Creatinine-, Cystatin C-, and Combined Creatinine-Cystatin C-Based Equations for Estimating Glomerular Filtration Rate: A Real-World Analysis in Patients with Chronic Kidney Disease. Int J Mol Sci. 2025;27(1). DOI: 10.3390/ijms27010364

10. Preda EC, Oprea O, Barabas ZA, Simion IP, Fotache AM, Dobreanu M. Impact of creatinine measurement methods on eGFR and GFR category assignment. J Crit Care Med (Targu Mures). 2026;12(2):199-206. DOI: 10.62838/jccm-2026-0007
11. Coucke W, China B, Delattre I, Lenga Y, Van Blerk M, Van Campenhout C, et al. Comparison of different approaches to evaluate External Quality Assessment Data. Clin Chim Acta. 2012;413(5-6):582-6. DOI: 10.1016/j.cca.2011.11.030
12. van der Hagen EAE, Weykamp C, Sandberg S, Stavelin AV, MacKenzie F, Miller WG. Feasibility for aggregation of commutable external quality assessment results to evaluate metrological traceability and agreement among results. Clin Chem Lab Med. 2020;59(1):117-25. DOI: 10.1515/cclm-2020-0736
13. Carobbi C. A modified ISO 13528 robust analysis (Algorithm A) that takes measurement uncertainty into account. Measurement. 2017;110:296-306. DOI: 10.1016/j.measurement.2017.07.006
14. Ilinca R, Ganea I. Making the most out of proficiency testing participation for a medical laboratory from the standardization point of view. Revista Romana de Medicina de Laborator. 2023;31(1):9-14. DOI: 10.2478/rrlm-2023-0003
15. Tsamatsoulis D. Comparing the Robustness of Statistical Estimators of Proficiency Testing Schemes for a Limited Number of Participants. Computation. 2022;10(3). DOI: 10.3390/computation10030044
16. Ilinca R, Chiriac IA, Luțescu DA, Ganea I, Hristodorescu-Grigore S, Dănciulescu-Miulescu R-E. Understanding the key differences between ISO 15189:2022 and ISO 15189:2012 for an improved medical laboratory quality of service. Revista Romana de Medicina de Laborator. 2023;31(2):77-82. DOI: 10.2478/rrlm-2023-0011
17. International Organization for Standardization. ISO/IEC 17011:2017. Conformity assessment — Requirements for accreditation bodies accrediting conformity assessment bodies. Standard. Geneva, Switzerland: International Organization for Standardization; 2017.
18. International Organization for Standardization. ISO 13528:2022 Statistical methods for use in proficiency testing by interlaboratory comparison. Standard. Geneva, Switzerland: International Organization for Standardization; 2022.
19. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: pros, cons and progress. Scand J Clin Lab Invest. 1999;59(7):491-500. DOI: 10.1080/00365519950185229
20. Petersen PH, Fraser CG. Desirable standards for laboratory tests if they are to fulfill medical needs. Clinical Chemistry. 1993;39(7):1447-55. DOI: 10.1093/clinchem/39.7.1447
21. Joint Committee for Guides in Metrology. JCGM 100:2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections). Guide. Sèvres, France: Joint Committee for Guides in Metrology; 2008.
22. Gerloff N, Burns CC. Evaluating the Effectiveness of External Molecular Proficiency Testing in the Global Polio Laboratory Network, 2021-2022. Pathogens. 2024;13(11). DOI: 10.3390/pathogens13111014
23. Oosterhuis WP, Bayat H, Armbruster D, Coskun A, Freeman KP, Kallner A, et al. The use of error and uncertainty methods in the medical laboratory. Clin Chem Lab Med. 2018;56(2):209-19. DOI: 10.1515/cclm-2017-0341
24. Sciacovelli L, Secchiero S, Padoan A, Plebani M. External quality assessment programs in the context of ISO 15189 accreditation. Clin Chem Lab Med. 2018;56(10):1644-54. DOI: 10.1515/cclm-2017-1179
25. Oprea OR, Preda EC, Mănescu IB, Dobreanu M. Setting up an own laboratory performance-based internal quality control plan - a model for complete blood count. Revista Romana de Medicina de Laborator. 2022;30(4):477-82. DOI: 10.2478/rrlm-2022-0036
26. International Organization for Standardization. ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories. Standard. Geneva, Switzerland: International Organization for Standardization; 2017.