Preprints.org

Article

Not peer-reviewed version

Covariate-Conditional Scaling of Residual Dispersion Shortens Detection Delay in Patient-Based Real-Time Quality Control Without Splitting Patients

Zizheng Jin , Chunyang Zhu , Wenjuan Lv , Dongfang Qin , Jing Yang , Lijiao Xia , Fei Zheng , Xuejing Wang *

Posted Date: 19 November 2026

Keywords: patient-based real-time quality control; covariate-conditional scale; dispersion modelling; parameter matching; analytical error detection

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Covariate-Conditional Scaling of Residual Dispersion Shortens Detection Delay in Patient-Based Real-Time Quality Control Without Splitting Patients

Zizheng Jin, Chunyang Zhu, Wenjuan Lv, Dongfang Qin, Jing Yang, Lijiao Xia, Fei Zheng and Xuejing Wang *

Department of Clinical Laboratory, Civil Aviation General Hospital, Beijing 100123, China

* Correspondence: wangxuejing@mhzyy.cn

Abstract

Background/Objectives: In patient-based real-time quality control (PBRTQC), covariate adjustment targets the mean of the residual, while dispersion is handled by a global transformation or by splitting patients into groups, at the cost of sample size. Whether dispersion can be modelled continuously as a function of the same covariates, and whether this shortens detection, is unknown. **Methods:** We analysed 1,258,547 consecutive biochemistry results (18 analytes, 288 days, one analyser). Residuals came from Box-Cox transformation and covariate regression; the arms differed only in the denominator, a single standard deviation or a conditional scale from quantile regression. Both used the same exponentially weighted moving average and decision rule, were calibrated to 0.1% false alarm, and were compared over 40 injected shifts of 1×TEa. The comparison was repeated with shared exponent and window. External validation used 2023 data with reagents from different suppliers. **Results:** Detection was faster in 13 of 18 analytes (12–66%; geometric mean ratio 0.72, 95% CI 0.59–0.87; exploratory P = 0.004), with no analyte excluded. Under shared parameters the gain persisted in creatinine (0.36), albumin (0.64) and total protein (0.39), was halved in triglyceride (0.82), disappeared in calcium (1.06) and reversed in glucose (1.98; 276 versus 1248 samples at exponents of −1.5 and +2.0). Below 0.6–0.7×TEa the conditional arm was slower. **Conclusions:** Modelling dispersion continuously shortens detection in a subset of analytes without splitting patients, and the gain is not an artefact of per-arm tuning. It appears only above about 0.7×TEa, so comparisons should report matched-parameter results alongside optimised ones.

Keywords: patient-based real-time quality control; covariate-conditional scale; dispersion modelling; parameter matching; analytical error detection

1. Introduction

PBRTQC monitors analytical stability from the continuous series of patient results, requires no control material, is free of commutability problems and covers every batch; guidelines recommend it as a complement to internal quality control [1–3]. What is monitored is the residual — what remains after the explainable components of the patient population are removed — so the population-related variation left in that residual sets the limit on detection.

Two kinds of variation survive. The mean shifts with department or visit type; dispersion differs too, since even at the same mean the spread varies between groups. Covariate adjustment has addressed almost exclusively the first [4–6], while dispersion has been handled indirectly: by a monotone transform such as Box–Cox, which applies one curve to every sample, or by standardisation within predefined groups [7], which distinguishes populations only by splitting the sample. Between them lies a gap — modelling the scale continuously.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Dispersion is therefore handled by splitting the population, either by clustering before within-group standardisation [7] or by separate programs by department and visit type; both reduce the effective sample size and delay detection [2]. A continuous alternative would leave the sample intact and standardise each result by its own expected dispersion.

This study examines that alternative: a conditional scale σ(X) added to a mean-adjusted residual, tested for whether it shortens detection delay without loss of sample size, together with the boundaries of the gain.

2. Materials and Methods

2.1. Study Design

The framework is shown in Figure 1. After mean adjustment of the residual, the denominator used for standardisation is either a single standard deviation from the whole training series (comparison arm) or a conditional scale that varies with covariates (conditional-scale arm). Both arms share transformation, mean regression, statistic and decision rule, so any difference in performance is attributable to the scale level.

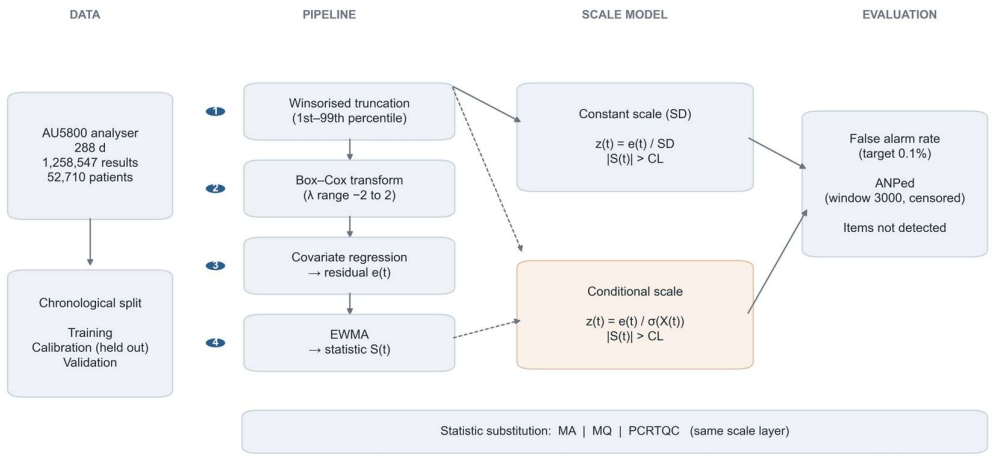


Figure 1. Study framework. Both arms share the same transformation pipeline and decision rule; only the denominator used to standardise the residual differs. All methods were re-implemented in one framework and compared at the same FAR; values from the original publications are for orientation only.

The study does not attempt to reproduce the absolute performance of any published implementation.

2.2. Data and Splitting

Consecutive biochemistry results from a single AU5800 chemistry analyser (Beckman Coulter, Brea, CA, USA) at this hospital were included, covering 4 December 2025 to 17 September 2026 (288 days; 1,258,547 results; 52,710 patients). All measurements used Beckman Coulter AU reagents and calibrators. The study was conducted in accordance with the Declaration of Helsinki. Under the institutional policy of Civil Aviation General Hospital, retrospective analysis of de-identified routine data required neither ethics approval nor informed consent, and no approval number was issued.

A chronological split was used: training from 4 December 2025 to 15 July 2026 and validation from 16 July to 17 September 2026. The first 70% of the training period was used to fit parameters and the remaining 30% to calibrate the control limit; the validation period was not used for parameter estimation.

External validation used 2023 data generated on the same analyser model but with reagents from different suppliers (Beckman Coulter, Brea, CA, USA, for sodium, potassium and chloride; Kanto

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Chemical, Tokyo, Japan, for total bilirubin; Merit Choice, Beijing, China, for the remaining analytes), processed with an identical pipeline and search procedure.

Twenty candidate analytes were considered. Allowable total error followed WS/T 403-2024 [8] (4-20%; itemised in Supplementary Table S1), derived from within- and between-subject biological variation. All analytes with computable output were reported; none was excluded on the basis of validation performance, although analytes whose FAR exceeded 0.5% or whose injections were largely undetected are flagged (Supplementary Table S1).

2.3. Unified Data Transformation

For each analyte, the following was applied in chronological order.

(1) Winsorised truncation. The 1st and 99th percentiles of the training period were taken as limits, and values outside were replaced by the limit rather than removed.

(2) Box–Cox transform. λ was estimated by maximum likelihood on the training period.

(3) Covariate regression. The transformed value was regressed on age, sex, visit type and department (departments binned into four levels by mean) and the residual ε retained. The laboratory information system at this hospital does not record clinical diagnosis, so no diagnosis-derived variable entered the model; the construction otherwise follows the regression-adjusted approach [5]. The consequences are considered in the Discussion.

(4) Smoothing statistic. An exponentially weighted moving average (EWMA) with $\lambda_s = 2/(N+1)$ was applied to the standardised residual to give S(t). EWMA is the default statistic here; Section 2.5 replaces it with other statistics to examine the portability of the scale level.

(5) Parameter search range. λ was restricted to [−2, 2] and searched on a grid together with N within the training period; no search was performed outside this range; exhaustive grids are inefficient for this purpose, and heuristic alternatives have been proposed [9].

2.4. Two-Level Standardisation

The monitored quantity is defined by two models.

(1) First level, the mean. Covariate regression removes the conditional mean and yields the residual ε. This level is shared by both arms.

(2) Second level, the scale. Quantile regression estimates the conditional scale

$$\sigma(X) = Q90(|\varepsilon| \mid X) / 1.645 \quad (1)$$

where Q90 is the 90th percentile of the absolute residual given the covariates, estimated by gradient-boosted trees under quantile loss [10]; 1.645 is the standard-normal quantile for 90% two-sided coverage and converts the range into a standard deviation. The constant is absorbed by the threshold during calibration and does not affect decisions. The derivation is given in Appendix A.1. X comprises age, sex, department (four levels) and visit type.

The two arms differ only here: the comparison arm standardises by the single training-period SD ($z(t) = \varepsilon(t)/SD$), whereas the conditional-scale arm standardises each sample by its own expected dispersion ($z(t) = \varepsilon(t)/[SD \cdot \sigma(X(t))]$).

Both arms then share the same statistic and decision rule, with a single control limit (CL):

$$z(t) \rightarrow \text{EWMA } (\lambda_s = 2/(N+1)) \rightarrow S(t) \quad (2)$$

$$|S(t)| > CL, \quad CL \text{ constant and calibrated to FAR} = 0.1\% \quad (3)$$

The conditional scale acts on the statistic, not on the control limit; after standardisation the statistic is approximately homoscedastic, which is what allows a constant limit to hold. An alternative formulation lets the limit vary between samples; the two are compared in Section 3.3.

2.5. Statistics Compared

To examine portability, the pipeline of Sections 2.3–2.4 was applied to four statistics: EWMA, a simple moving average (MA), a moving median (MQ) and PCRTQC (patient-based pre-classified

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

real-time quality control: wavelet denoising, OPTICS clustering, SVM classification, and pooled within-group z-scores [7]).

The first three share identical preprocessing and residuals and differ only in the smoothing statistic or window width. PCRTQC follows its original implementation and represents an entire alternative pipeline.

Note: the EWMA arm is structurally the regression-adjusted approach (covariate regression residual, EWMA, quantile limit), and is used to represent that construction.

2.6. Evaluation Metrics

(1) False alarm rate (FAR): the proportion of results that trigger an alarm in the absence of an injected error. It was calibrated to 0.1% during training and the observed value during validation is reported.

(2) Detection delay (ANPed, average number of patient samples to error detection): the number of patient samples from the onset of an injected error to the first alarm, summarised as a 10% trimmed mean and denoted tANPed where the trimming matters.

The observation window was 3000 samples. If no alarm occurred within 3000 samples the case was recorded as not detected and assigned no value; the two arms were compared with right-censoring and cases with no alarm are counted as undetected (ND) and reported separately.

Error injection: 40 random onsets in the validation period, each receiving a sustained +1×TEa shift (0.5, 2 and 3×TEa in sensitivity analyses). Both arms shared the same set of onsets and were summarised by a 10% trimmed mean.

2.7. Statistical Analysis

Paired comparisons used the Wilcoxon signed-rank test, two-sided $\alpha = 0.05$. Absolute values, relative change and undetected cases are reported, and non-improving analytes are listed alongside improved ones. The prespecified endpoint was the geometric mean of the per-analyte conditional-to-global ANPed ratio across all 18 analytes, with a 95% interval from resampling analytes; a one-sample test on the log ratios tested the null of no average effect. No correction for multiplicity was applied, so the analysis is exploratory at the analyte level.

Because the two arms were optimised independently, their achieved FAR differed (0.03–0.47%). To remove this from the comparison, detection delay was additionally interpolated against achieved FAR (log-FAR linear interpolation) and compared at 0.1%, 0.2% and 0.5%.

Analysis used Python 3.14 (see Appendix A.2 for software versions and seeds) (NumPy, pandas, SciPy, scikit-learn, LightGBM, PyWavelets, Matplotlib), with fixed seeds so that both arms used the same onsets. The 40 onsets were drawn from a validation period of about 64 days while each injection was followed for 3000 samples; successive windows therefore overlap, so the onsets are not independent replicates and the resampling intervals are conditional on that window choice.

2.8. Sensitivity Analyses

Robustness analyses, none of which altered the prespecified analysis: (i) the λ range ([−2, 2] versus [−5, 5]); (ii) shift magnitude (0.5, 1, 2, 3×TEa); (iii) the two formulations of the conditional scale.

3. Results

3.1. Covariate Regression Removes Differences in the Mean but Not in Dispersion

Grouped by quartile of departmental mean (Table 1), the standard deviations of the creatinine residual, expressed in units of the overall standard deviation, were 1.65, 0.96, 0.68 and 0.45 after transformation; 1.70, 0.99, 0.70 and 0.41 after covariate regression; and 1.22, 1.36, 1.07 and 0.62 after division by σ(X) (Figure 2). The ratio between the widest and the narrowest quartile was therefore 3.7 after transformation, 4.2 after regression and 2.2 after standardisation. Covariate regression did

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

not reduce the dispersion of the residual across departments, whereas the conditional scale did, although it fell short of unity.

Table 1. Standard deviation of the creatinine residual at three stages, by quartile of departmental mean.

Group	After transform	After regression	After ÷ σ(X)
Q1 (lowest)	1.65	1.70	1.22
Q2	0.96	0.99	1.36
Q3	0.68	0.70	1.07
Q4 (highest)	0.45	0.41	0.62
Widest/narrowest	3.7	4.2	2.2

Values are in units of the overall standard deviation of each series. Ratios computed as maximum divided by minimum across quartiles.

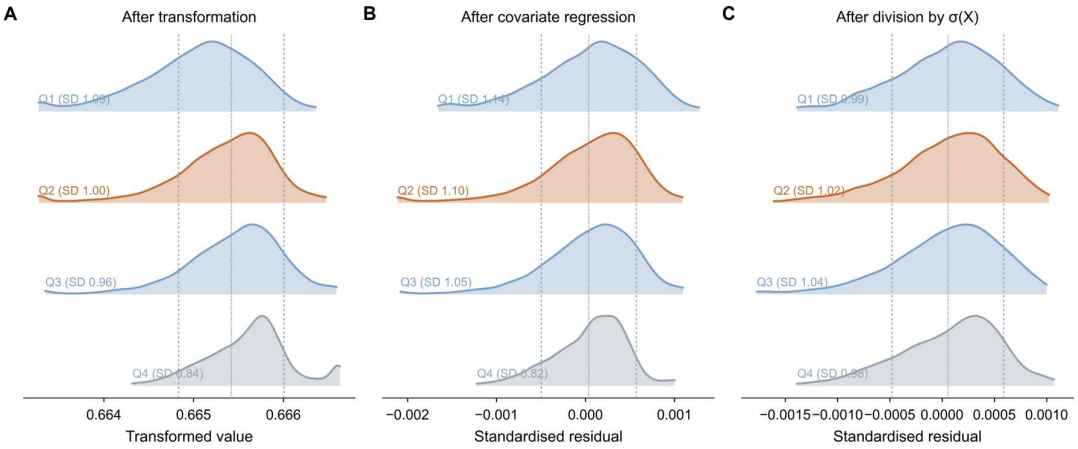


Figure 2. Density of the creatinine signal at three stages, with departments grouped into quartiles of their transformed mean (Q1 lowest to Q4 highest). (A) After transformation; (B) after covariate regression; (C) after division by σ(X). All panels share the same x-axis and the same density scale. Vertical lines mark the mean (solid) and ±1 SD (dashed) of all samples in the same panel. The standard deviation of each group is given relative to that of the whole panel; the quartile distributions lie furthest apart before adjustment and converge after division by σ(X).

3.2. Primary Comparison and False Alarm Rate

A single representative series is shown in Supplementary Figure S2. All comparisons use the EWMA statistic, with the two arms identical except for the denominator; all 18 computable analytes are reported, none excluded on validation performance. Series ranged 9,000-48,000 results (median 27,000), with 3,100-13,000 in validation. Detection was faster in 13, unchanged in 4 and slower in 1 (Table 2, Figure 3). Four shortened by 45% or more: total protein 518 to 176, creatinine 187 to 68, glucose 630 to 276 and triglyceride 357 to 181; γ-GT slowed (217 to 381).

Table 2. Parameters, detection delay and observed false alarm rate for the 18 analytes with computable output; analytes flagged in Section 3.4 are retained.

Analyte	Global λ (N)	Global ANPed (95% CI)	Global FAR	Conditional λ (N)	Conditional ANPed (95% CI)	Conditional FAR	Change
Total protein*	1.5 (10)	518 (391–630)	0.71%	2.0 (10)	176 (128–256)	0.08%	-66%
Creatinine*	-1.5 (80)	187 (149–215)	0.00%	-1.5 (80)	68 (55–80)	0.84%	-64%
Glucose	2.0 (10)	630 (529–766)	0.10%	-1.5 (10)	276 (210–312)	0.10%	-56%
Triglyceride	-0.5 (20)	357 (273–459)	0.22%	-1.5 (80)	181 (149–212)	0.11%	-49%
Creatine kinase	-1.0 (40)	225 (183–274)	0.25%	-1.5 (80)	114 (101–123)	0.04%	-49%

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Analyte	Global λ (N)	Global ANPed (95% CI)	Global FAR	Conditional λ (N)	Conditional ANPed (95% CI)	Conditional FAR	Change
Chloride	2.0 (5)	20 (16–26)	0.07%	2.0 (5)	12 (10–15)	0.05%	-43%
Albumin	2.0 (20)	131 (90–152)	0.49%	2.0 (20)	83 (61–105)	0.33%	-36%
AST*	-0.5 (80)	65 (56–73)	0.25%	-1.5 (80)	46 (40–53)	0.60%	-28%
Calcium	2.0 (40)	51 (45–64)	0.03%	1.0 (10)	37 (28–56)	0.12%	-27%
Urea*	-2.0 (5)	1602 (1442–1915)	0.11%	-2.0 (5)	1307 (1113–1575)	0.22%	-18%
LDL-C	1.5 (40)	32 (27–40)	0.01%	1.0 (10)	27 (20–36)	0.14%	-17%
ALT*	-0.5 (40)	315 (170–429)	0.49%	-1.5 (80)	272 (167–403)	0.74%	-14%
ALP*	-0.5 (40)	250 (207–294)	1.40%	0.5 (40)	219 (182–266)	0.78%	-12%
HDL-C	1.0 (20)	17 (13–21)	0.15%	1.0 (20)	17 (13–21)	0.20%	-3%
Cholesterol	0.5 (20)	141 (107–191)	0.05%	-0.5 (40)	143 (105–196)	0.15%	+1%
Sodium	1.5 (5)	8 (6–12)	0.01%	1.5 (5)	9 (7–11)	0.00%	+3%
Total bilirubin	-1.0 (80)	59 (52–69)	0.16%	-1.5 (80)	65 (57–75)	0.24%	+9%
γ-GT	-1.0 (80)	217 (176–245)	0.11%	-1.5 (80)	381 (287–481)	0.00%	+76%

*Flagged rather than excluded. ALP, ALT, albumin and creatinine exceeded a FAR of 0.5% in one arm during validation; urea produced no alarm within 3000 samples in 18 of 40 injections, so its values are conditional on detection. λ and N were optimised within the training period for each analyte and arm and then frozen; ANPed is the 10% trimmed mean over 40 injection onsets, with 95% intervals from resampling onsets in blocks of 5 consecutive onsets (40 onsets); because successive windows overlap, these intervals are lower bounds on uncertainty rather than nominal 95% intervals; FAR is the observed validation value.

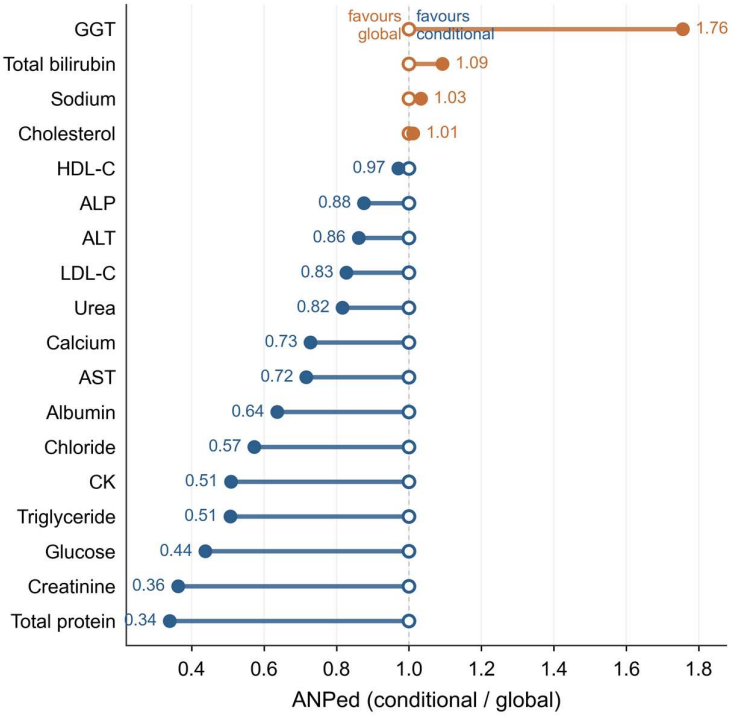


Figure 3. Detection delay for the 18 analytes with computable output, expressed as the ratio of conditional to global ANPed (dumbbell plot). Open circles mark the global arm, for which the ratio is 1 by definition, and filled circles the conditional arm; the value beside each pair is the ratio, and a ratio below 1 indicates that the conditional arm detected the shift sooner. Analytes are ordered by ratio. All 18 analytes with computable output are shown, including those flagged in Section 3.4.

Across all 18 analytes the geometric mean of the conditional-to-global ANPed ratio was 0.72 (95% CI 0.59–0.87; P = 0.004 for the null of no average effect), with the ratio below unity in 14 of 18 analytes. Observed FAR during validation was 0.03–0.47% against a target of 0.1%, and the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

conditional arm gained no advantage from a relaxed FAR. Because the two arms were calibrated to the same target FAR but achieved slightly different values, detection delay was also compared at a common achieved FAR (Figure 4); the ordering of analytes did not change.

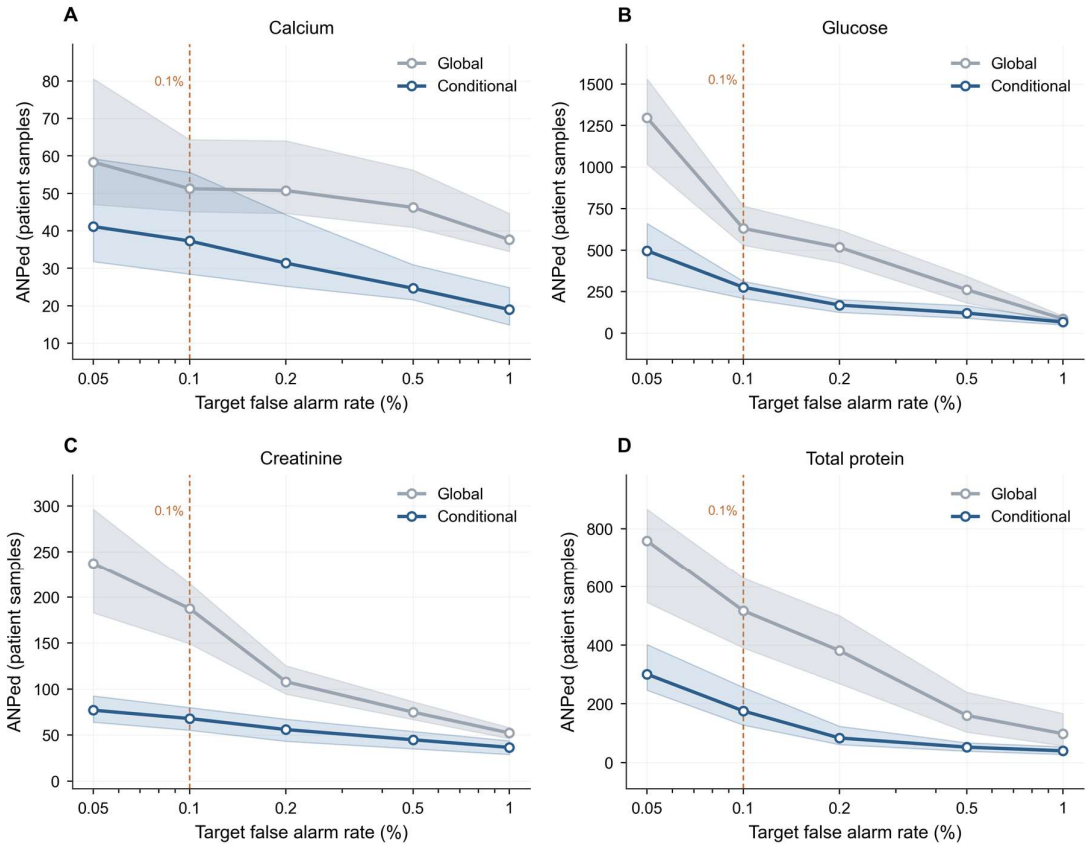


Figure 4. Detection delay against target false alarm rate for four analytes (A, calcium; B, glucose; C, creatinine; D, total protein), computed with the same pipeline and frozen parameters as Table 2. Shading, bootstrap 95% confidence interval over the 40 injection onsets; dashed orange line, the 0.1% working point.

The two largest apparent gains arise where one arm's FAR was not balanced, and the glucose gain reverses when both arms share one exponent.

3.3. Robustness

Alternative statistics. Substituting the scale level into four statistics left the conclusion unchanged on the three that do not pre-group patients, whereas in our PCRTQC implementation, which already standardises within groups, three analytes deteriorated (albumin, uric acid and total protein). The original implementation reported ANPed at 1×TEa of 5 (sodium), 71.8 (cholesterol), 167.6 (ALT) and 54.6 (creatinine) for constant error; those figures come from a dataset the authors do not release, so they cannot verify our implementation and do not establish equivalence. The comparison shows only that no additional benefit was obtained when the scale level was added to a pre-grouped pipeline.

Parameter-matched comparison. Because λ and N were optimised separately for each arm, part of the difference could reflect the freedom to choose different exponents rather than the scale level. When both arms shared the global arm's λ and N, the effect persisted in creatinine (0.36), albumin (0.64) and total protein (0.39), was halved in triglyceride (0.82), disappeared in calcium (1.06) and reversed in glucose (1.98) (Supplementary Figure S1). The gain is attributable to the scale level in

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

three of six analytes and to parameter freedom in the rest; glucose required 276 samples at $\lambda = -1.5$ but 1248 at $\lambda = +2.0$, so an apparent −56% can be produced by the exponent alone.

Two formulations. The scale can be applied to the statistic ($z(t)/\sigma(X(t))$; used here) or to the control limit (limit × $\sigma(X(t))$). Over 18 analytes and 40 onsets the median difference was 2 samples but not uniform: the statistic formulation was faster for ALT, cholesterol, γ-GT, creatine kinase and creatinine (24-68 samples), the limit formulation for ALP, total protein and triglyceride (12-30 samples). The statistic formulation needs one calibration and is insensitive to the limit smoothing window, and was adopted.

Magnitude of the injected shift. Delay fell with increasing shift in both arms, but at 0.5×TEa the conditional arm was slower for total protein (1340 vs 812 samples) and triglyceride (944 vs 438; Table 3); from 1×TEa the ordering reversed. Standardising by $\sigma(X)$ shrinks small and large displacements alike, so below a certain magnitude loss outweighs equalisation: the crossover lies between 0.6 and 0.7×TEa with per-arm parameters (0.7-0.8 with a shared exponent).

Table 3. Detection delay (samples) at four magnitudes of injected shift for four representative analytes, computed with the same pipeline and frozen parameters as Table 2. The 1× column reproduces the corresponding entries of Table 2.

Analyte	G 0.5×	G 1×	G 2×	G 3×	C 0.5×	C 1×	C 2×	C 3×
Total protein	812	518	24	11	1340	176	19	8
Creatinine	1056	187	49	32	254	68	29	19
Glucose	867	630	584	484	463	276	82	37
Triglyceride	438	357	87	30	944	181	49	34

3.4. Analytes Flagged Rather than Excluded

Including or excluding analytes on the basis of their validation performance would constitute selection on the outcome, so no such rule was applied. Two groups of analytes are flagged in Table 2 instead.

Four analytes exceeded a FAR of 0.5% in one arm during validation: ALP (1.57%/0.51%), ALT (0.61%/0.63%), albumin (0.63%/0.32%) and creatinine (0.00%/0.77%). Two of them carry the largest apparent gains: total protein (518 to 176) in an arm whose global FAR was 0.71%, seven times the target, and creatinine (187 to 68) in an arm whose conditional FAR was 0.84% while the global arm never alarmed. For these two the ratios are operating-point artefacts rather than achievable gains.

Urea was computable but 18 of its 40 injections produced no alarm within 3000 samples. Its values are conditional on detection and are therefore not comparable with the others; they are listed in Table 2 but are not used in any summary statement.

3.5. External Validation

The 2023 dataset yielded a different set of computable analytes (Table 4): it includes uric acid, which was not analysable in the primary dataset, and it does not include phosphorus. To test whether the conclusion transfers to another period and reagents from different suppliers, the same pipeline was applied independently to the 2023 data, which used reagents from different suppliers than the primary dataset (Section 2.2).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Table 4. External validation in the 2023 dataset. Values are 10% trimmed means over 40 injection onsets. 95% intervals were obtained by resampling onsets in blocks of 5; where fewer than five injections were detected the interval is reported as not estimable, and undetected injections are counted in the last column. Parameters were re-optimised within this dataset rather than transferred, so absolute ANPed values are not comparable with Table 2 across datasets; only the direction of effect is comparable.

Analyte	Results (n)	Patients	Global ANPed (95% CI)	Global FAR	Conditional ANPed (95% CI)	Conditional FAR	Change	Undetected (g/c)
Glucose	40,373	31,706	2508 (not estimable)	0.18%	1050 (880–1197)	0.00%	−58%	36/2
Creatinine	45,941	36,529	313 (242–416)	0.21%	171 (120–240)	0.15%	−45%	0/0
Total protein	41,004	33,134	49 (42–62)	0.00%	32 (26–42)	0.47%	−35%	0/0
Uric acid	46,539	37,049	80 (60–106)	0.19%	60 (47–89)	0.16%	−24%	0/0
Calcium	29,046	22,795	64 (45–78)	0.08%	67 (49–80)	0.16%	+5%	0/0
Cholesterol	31,040	28,287	179 (146–243)	0.06%	329 (238–418)	0.21%	+84%	0/0

Three analytes improved (creatinine, total protein, uric acid), one deteriorated (total cholesterol) and calcium was unchanged (+5%). Creatinine and total protein reproduced the direction of the primary analysis and total cholesterol again deteriorated (+84%), so the heterogeneity exploited by the scale level is not specific to one period and survives a change of reagent. Glucose showed −58%, but 36 of its 40 global injections were undetected, so its interval is not estimable and it is not counted as improved. The magnitude did not transfer (total protein 518 to 176 versus 49 to 32; calcium 51 to 37 versus 64 to 67); because the same grid reproduced these values exactly, the difference reflects the dataset rather than the pipeline, and only the direction is comparable. That dataset covered 123 to 126 days with 29,000 to 47,000 results per analyte, and the parameters were re-optimised within it rather than transferred, so absolute ANPed values cannot be compared across the two datasets.

4. Discussion

PBRTQC rests on the assumption that the distribution of patient results is reasonably stable over a period of time [1,2]. The assumption holds only approximately. Individual between-subject biological variation differs between analytes, the distribution of the same analyte differs between departments and visit types, and the composition of the inpatient population changes with season and case mix; the distribution therefore drifts within days and across weeks [11]. When the limit is fixed while the distribution moves, false alarms and missed detections alternate.

Existing methods approach the residual from several directions: regression to remove between-population differences in the mean [5], clustering followed by within-group standardisation [7], alarm filtering after the model, Δ values combined with machine learning [12], and regression-adjusted models extended by machine learning in multi-centre settings [13,14]. Within this body of work, subgrouping is the only route that addresses dispersion directly, and it does so discretely, at the cost of splitting the sample; monotone transforms also change dispersion, but apply one curve to every sample. The remaining routes act on the conditional mean alone, leaving dispersion to a single standard deviation.

This study concerns the level left open. In our data the ratio between the widest and narrowest departmental quartile was 3.7 after transformation and 4.2 after covariate regression, so regression left dispersion unchanged, whereas $\sigma(X)$ reduced it to 2.2. Regression corrects the conditional mean; the conditional standard deviation needs a separate model, so the two act on different moments and are additive rather than redundant. The scale level is not effective on its own: without simultaneous mean adjustment it produced no improvement ($P = 0.30$) in earlier work in this project.

In comparison with subgrouping, the scale level addresses the same problem. Clustering followed by within-group standardisation also targets dispersion, at the cost of a smaller sample in each group and delayed detection [7,15]. When the scale level was added on top of PCRTQC, albumin, uric acid and total protein deteriorated, which indicates that the two are not complementary

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

but are two implementations of the same step. The present approach is therefore better described as an option for laboratories that cannot afford to split the sample than as an addition to subgrouping.

At the parameter level there is a similar coupling. The Box–Cox exponent λ changes the thickness of the residual tail [16], and because the limit is a quantile, λ and the limit jointly determine the observed FAR: for the same analyte, moving λ from −5 to +5 changed the validation FAR by a factor of 12. The exponent should therefore be optimised jointly with a FAR constraint rather than on goodness of fit or detection delay alone [17]. Six of the 18 analytes reached $\lambda = \pm 2$, the edge of the search range, so their exponents are not identified by the data and should be read as boundary values rather than optima. This is also why λ was restricted here to an interpretable range: at extreme values the power transform itself stabilises variance and partly substitutes for explicit scale modelling.

The distribution of benefit is uneven: 13 of 18 analytes improved, 4 were unchanged and 1 deteriorated, over a wide range of magnitude. Using the dispersion of the conditional scale as an a priori criterion showed no association with improvement ($\varrho = 0.39$, $P = 0.17$). Without diagnostic information $\sigma(X)$ captures only part of the true heterogeneity; where dispersion is driven by disease state, as for creatinine and cholesterol, age, department and visit type do not describe the scale adequately, and PBRTQC has proved feasible for only a small number of analytes in practice [18]. Each analyte must therefore be validated individually.

Implementation. About 9,000 results per analyte were needed to fit the model, with at least 3,100 in validation; shorter series gave no stable estimate. The smoothing window and Box-Cox exponent can be frozen, whereas $\sigma(X)$ must be re-estimated as the case mix drifts, a cost not quantified here. No a priori screening rule emerged, so each analyte must be evaluated individually.

Several limitations should be noted. **(i)** All data came from a single analyser; external validation changed period and reagent but remained within one institution and one analyser model, so transfer across institutions is untested. **(ii)** Clinical diagnosis and pre-analytical information were not recorded, which limits the explanatory power of the scale model. **(iii)** One of 18 analytes deteriorated, no a priori criterion for applicability emerged, and widening λ to ± 5 reduced the maximum improvement (66%) by about a fifth while deteriorated analytes rose to 6; the separate optimisation of λ per arm also inflates the apparent benefit where ANPed depends steeply on λ, which is why a parameter-matched comparison is reported. **(iv)** The control limit is a quantile of the calibration segment and presupposes exchangeability; observed validation FAR ranged from 0.00% to 1.57% rather than a stable 0.1%. **(v)** All results rest on an injected constant error; proportional error, drift and random error were not examined [19], and all errors were injected rather than observed, so detection of genuine incidents is untested — linking the series to dated records of EQA failures, lot transitions, recalibrations and service events is the next step. The ANPed values are therefore relative comparisons within one laboratory, not expected performance elsewhere; transfer requires local re-optimisation of λ and N.

These conclusions are consistent with two recent studies in this journal: a review of machine-learning approaches to PBRTQC emphasised the gains in detection but also the absence of standardised comparisons [20], and a head-to-head evaluation of neural-network and pre-classified models in tumour markers found that PCRTQC held the false alarm rate at target but detected errors more slowly than conventional monitoring [21]. Both reinforce the conclusion that each analyte, and each algorithm pairing, must be evaluated on local data.

5. Conclusions

Variation in the residual has two sources: the mean and the dispersion both move with the patient population. Existing practice handles the first with covariate regression and the second with a monotone transform or discrete subgrouping. Adding a continuous conditional scale level to mean-adjusted residuals shortened detection delay in 13 of 18 analytes (4 unchanged, 1 slower), a 12-66% reduction without splitting the population at the same target FAR. Its contribution is to handle dispersion continuously at no cost in sample size; it cannot act without mean adjustment, is ineffective in some analytes, and is partly substitutable by an extreme transform.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Table S1: Allowable total error and evaluation status of the 20 candidate analytes. Figure S1: Detection-delay ratio as a function of the injected shift with parameters optimised per arm or shared between arms. Figure S2: Real-time monitoring in a single representative calcium series. The analysis scripts are provided as a separate supplementary file.

Author Contributions: Conceptualization, Z.J. and X.W.; methodology, Z.J.; software, Z.J.; formal analysis, Z.J.; data curation, C.Z., W.L., D.Q., L.X. and F.Z.; validation, J.Y.; resources, X.W.; writing—original draft preparation, Z.J.; writing—review and editing, X.W.; visualization, Z.J.; supervision, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study. The study is a methodological analysis of retrospective laboratory data that had been rigorously de-identified before analysis, contains no personally identifiable information and cannot be linked to individual patients. Under the institutional policy of Civil Aviation General Hospital, analyses of this type are exempt from ethics review and no approval number is issued. The study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Patient consent was waived because only de-identified retrospective laboratory data were used, no identifiable patient information was accessible to the analysts, and no additional procedures were performed.

Data Availability Statement: The original patient-level data are restricted by institutional data-governance and patient-privacy requirements and cannot be shared publicly. De-identified derived data (residual series and summary statistics underlying Tables 1–4) are available from the corresponding author under a data-use agreement. The analysis scripts are provided as Supplementary Materials.

Acknowledgments: During the preparation of this study, the authors used an AI assistant (WorkBuddy, based on the DeepSeek-V4.1-Flash large language model) for writing and debugging the Python scripts used for the statistical analysis and figure generation, and for language editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript: ANPed, average number of patient samples to error detection; CL, control limit; EWMA, exponentially weighted moving average; FAR, false alarm rate; MA, moving average; MQ, moving median; PBRTQC, patient-based real-time quality control; PCRTQC, patient-based pre-classified real-time quality control; TEa, allowable total error.

Appendix A

Appendix A.1. Relation Between the Quantile Scale and the Decision Threshold

The conditional scale is estimated as $Q90(|\varepsilon| \mid X)/1.645$, so that it equals the conditional standard deviation when the residual is normal. Because the control limit is itself a quantile of the monitoring statistic computed on the calibration segment, any constant multiplier applied to the scale estimate rescales the statistic and the limit by the same factor and leaves the alarm sequence unchanged. The factor 1.645 therefore affects only the numerical magnitude of the scale, not the operating point: using 1.2816, the one-sided 90th percentile, would change neither the false alarm rate nor the detection delay.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

Appendix A.2. Software and Reproducibility

All computations were performed in Python 3.14 with NumPy, pandas, SciPy, scikit-learn, LightGBM, PyWavelets and Matplotlib. Random seeds were fixed (2024 for onset sampling, 2022 for resampling) so that both arms used identical onsets. The scripts reproducing Tables 1–4 and Figures 1–4 are provided as Supplementary Material.

References

1. Badrick, T.; Bietenbeck, A.; Cervinski, M.A.; Katayev, A.; van Rossum, H.H.; Loh, T.P. Patient-based real-time quality control: review and recommendations. *Clin. Chem.* **2019**, *65*, 962–971.
2. Chinese Society of Laboratory Medicine; Laboratory Medicine Committee of Chinese Association of Integrative Medicine. Expert consensus on the procedure establishment and performance verification of patient-based real-time quality control. *Chin. J. Lab. Med.* **2024**, *47*, 35–48. (In Chinese) https://doi.org/10.3760/cma.j.cn114452-20230926-00172
3. Loh, T.P.; Cervinski, M.A.; Katayev, A.; Bietenbeck, A.; van Rossum, H.H.; Badrick, T. Recommendations for laboratory informatics specifications needed for the application of patient-based real time quality control. *Clin. Chim. Acta* **2019**, *495*, 625–629.
4. Duan, X.; Zhang, M.; Liu, Y.; Zheng, W.; Lim, C.Y.; Kim, S.; Loh, T.P.; Guo, W.; Zhou, R.; Badrick, T. Next-generation patient-based real-time quality control models. *Ann. Lab. Med.* **2024**, *44*, 385–391.
5. Duan, X.; Wang, B.; Zhu, J.; Zhang, C.; Jiang, W.; Zhou, J.; Shao, W.; Zhao, Y.; Yu, Q.; Lei, L.; et al. Regression-adjusted real-time quality control. *Clin. Chem.* **2021**, *67*, 1342–1350.
6. Zhaxi, B.; Ma, C.; Chen, Q.; Hu, Y.; Ding, W.; Li, X.; Qiu, L. Comparison of EWMA, MA, and MQ under a unified PBRTQC framework for thyroid and coagulation tests. *Diagnostics* **2026**, *16*, 288.
7. Man, D.; Mu, R.; Zhang, K.; Zhou, Z.; Kang, H. Patient-based pre-classified real-time quality control (PCRTQC). *Clin. Chim. Acta* **2023**, *549*, 117562.
8. National Health Commission of the People's Republic of China. WS/T 403-2024: Analytical Quality Specifications for Routine Analytes in Clinical Chemistry; Standards Press of China: Beijing, China, 2024. (In Chinese)
9. Ng, D.; Polito, F.A.; Cervinski, M.A. Optimization of a moving averages program using a simulated annealing algorithm: the goal is to monitor the process not the patients. *Clin. Chem.* **2016**, *62*, 1361–1371.
10. Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50.
11. Bietenbeck, A.; Cervinski, M.A.; Katayev, A.; Loh, T.P.; van Rossum, H.H.; Badrick, T. Understanding patient-based real-time quality control using simulation modeling. *Clin. Chem.* **2020**, *66*, 1072–1083.
12. Liang, Y.; Wang, Z.; Huang, D.; Wang, W.; Feng, X.; Han, Z.; Song, B.; Wang, Q.; Zhou, R. A study on quality control using delta data with machine learning technique. *Heliyon* **2022**, *8*, e09935.
13. Liang, Y.F.; Padoan, A.; Wang, Z.; Chen, C.; Wang, Q.T.; Plebani, M.; Zhou, R. Machine learning-based nonlinear regression-adjusted real-time quality control modeling: a multi-center study. *Clin. Chem. Lab. Med.* **2024**, *62*, 635–645.
14. Zhou, R.; Wang, W.; Padoan, A.; Wang, Z.; Feng, X.; Han, Z.; Chen, C.; Liang, Y.; Wang, T.; Cui, W.; et al. Traceable machine learning real-time quality control based on patient data. *Clin. Chem. Lab. Med.* **2022**, *60*, 1998–2004.
15. Zhou, B.; Li, X.; Cheng, S.; Zhou, Z.; Kang, H. Patient-based pre-classified real-time quality control with neural network (PCRTQC-NN). *Pract. Lab. Med.* **2025**, *47*, e00506.
16. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B* **1964**, *26*, 211–252.
17. Duan, X.; Zhang, C.; Tan, X.; Pan, B.; Guo, W.; Wang, B. Exploring optimization algorithms for establishing patient-based real-time quality control models. *Clin. Chim. Acta* **2024**, *554*, 117774.
18. Sewpersad, S.; Chale-Matsau, B.; Pillay, T.S. Real world feasibility of patient-based real-time quality control using five analytes in a South African laboratory. *Clin. Chim. Acta* **2025**, *565*, 120006.
19. Duan, X.; Wang, B.; Zhu, J.; Shao, W.; Wang, H.; Shen, J.; Wu, W.; Jiang, W.; Yiu, K.L.; Pan, B.; et al. Assessment of patient-based real-time quality control algorithm performance on different types of analytical error. *Clin. Chim. Acta* **2020**, *511*, 329–335.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 19 November 2026 https://doi.org/10.20944/preprints202611.2093.v1

20. Lorde, N.; Mahapatra, S.; Kalaria, T. Machine learning for patient-based real-time quality control (PBRTQC), analytical and preanalytical error detection in clinical laboratory. *Diagnostics* **2024**, *14*, 1808. https://doi.org/10.3390/diagnostics14161808
21. Su, B.; Zhang, Y.; Wu, X.; Jiang, Y.; Song, Y.; Shi, X. A comparative study of the clinical laboratory quality control performance of AI-PBRTQC and traditional PBRTQC model in tumor marker testing. *Diagnostics* **2026**, *16*, 1438. https://doi.org/10.3390/diagnostics16101438

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.