Dean&Francis

ISSN 2959-409X

AI-Driven Patient-Based Real-Time Quality Control System Optimization And Bias Alarming Analysis

Zihan Liu

University of Washington, United States
Corresponding author: zihanliu@uw.edu

Abstract:

Patient-based real-time quality control (PBRTQC) is widely used in clinical laboratories, providing a sophisticated approach to monitoring the analytical performance of laboratory instruments using the results generated from actual patient samples. Compared with the traditional quality control (QC), which easily triggers false alarms, delayed error detection, and limited specificity and sensitivity, PBRTQC overcomes critical gaps by proficiently adopting machine learning. Implemented under a set of standardized rules and methods, PBRTQC enables intelligent error detection, handling of complex scenarios, and competitive operational efficiency gains. However, PBRTQC is still facing some challenges due to an imbalance of data resources and a limited ability to identify real-world issues when models are only trained on simulated datasets. This article discusses the development of PBRTQC from the initial Westgard rule to several typical AI-driven machine learning models, including comparing their mechanisms and the disparity in accuracy performance. The compelling advantages of the machine learning models will be highlighted, such as their highly precise model structure and algorithm. At the same time, current limitations of ML models and critical thinking for future projection are also discussed.

Keywords: PBRTQC, Machine Learning, quality control, Westgard Rules

1. Introduction

The core concept of traditional quality control(QC) is based on statistical process control (SPC), which uses tools to monitor, control, and improve a process. SPC helps to identify and minimize the variation in real-time to ensure a stable process. With the collected data for several measurements, the mean value and standard deviation can be calculated. Applying characterized Westgard rules to draw Levey-Jennings plots and then to estimate whether it is within the control range. The control rules can be 12s or 13s,

which means if the single control measurement exceeds a 2s range, the run is rejected.

With the innovation of Patient-based real-time quality control, the limitations of traditional QC can be effectively mitigated, such as false alarms, poor sensitivity to small shifts, and manual workload. PBRTQC uses real patient sample results to monitor instrument performance in real time, originating from a simple moving average(MA) that is used to reduce noise and stabilize trends to a more advanced EWMA model that aims to detect small shifts based on its flexible calculation more weight to recent data.

Machine learning plays an important role in the innovation of quality control. It adopts developed algorithms like isolation forest to remove abnormal patterns and random forest to perform efficient classification. More recently, automated convolutional neural networks (CNNs) have attracted great attention due to their extraordinary self-learning ability. Compared to the previous model, which needs to add the patient's age, sex, and several factors that will affect the results, CNN is good at learning and summarizing the existing patterns. The CSLR model, which applies stepwise linear regression, has achieved great success in improving sensitivity and early warning capability by introducing simulated biases into training data. Compared to the traditional model, where up to 39% of errors are missed, the CSLR model can detect 98% of all simulated errors.

Integrating AI models into PBRTQC offers advantages, including faster error detection, handling nonlinear, high-dimensional, and noisy data. AI models also achieved a breakthrough for adaptive thresholds, root cause analysis, and clinical integration.

However, these cutting-edge machine learning systems are facing some challenges that limit their development. AI models are over-reliant on simulated data during training, which might cause some bias to be embedded into the program. The trade-offs between sensitivity and specificity with wide vs. narrow QC limits still need further modification.

2. Principle Rule

James Westgard, the founder of the Westgard rule, built a set of multirole quality control, which uses a combination of decision standards to decide whether the analytical run is in-control or out-of-control in the 1970s as shown in figure 1[1]. The Levey-Jennings chart is always used to organize the data and give clear visual information. The correlated rule served as a reference to guide the final determination. Based on different conditions, the control limits are set as the mean plus or minus a certain time.

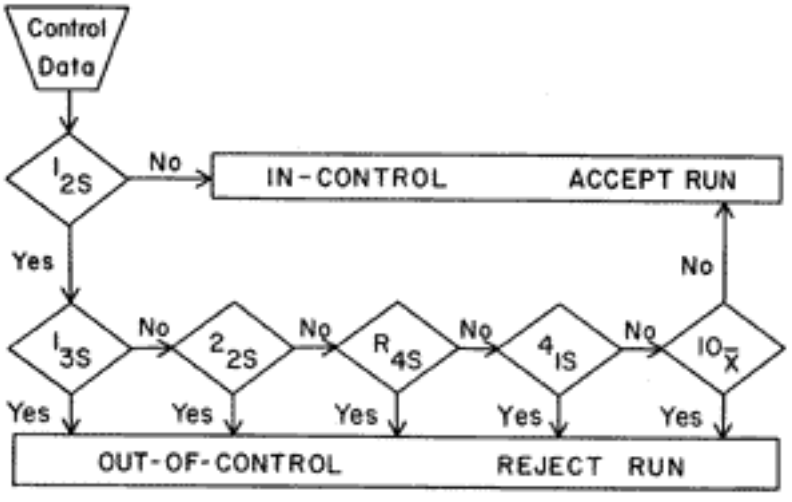

Figure 1: A combination of decision standards to decide whether the analytical run is in-control or out-od-control.

For example, 12s is a common warning sign within the control limit, used with the chart to show the run has exceeded the mean plus/minus 2s. This reminds further careful inspection should be carried out. 2 2s violation means that when 2 consecutive control measurements exceed the same mean plus/minus 2s, the run is rejected. R 4s means when 1 control measurement in a group exceeds the mean plus 2s and another exceeds the mean minus 2s. This method is easy to figure out if the run can be accepted without technical barriers. Figure 2 illustrates a run sample of 8x violation, and this output will guide further careful monitoring.

ISSN 2959-409X

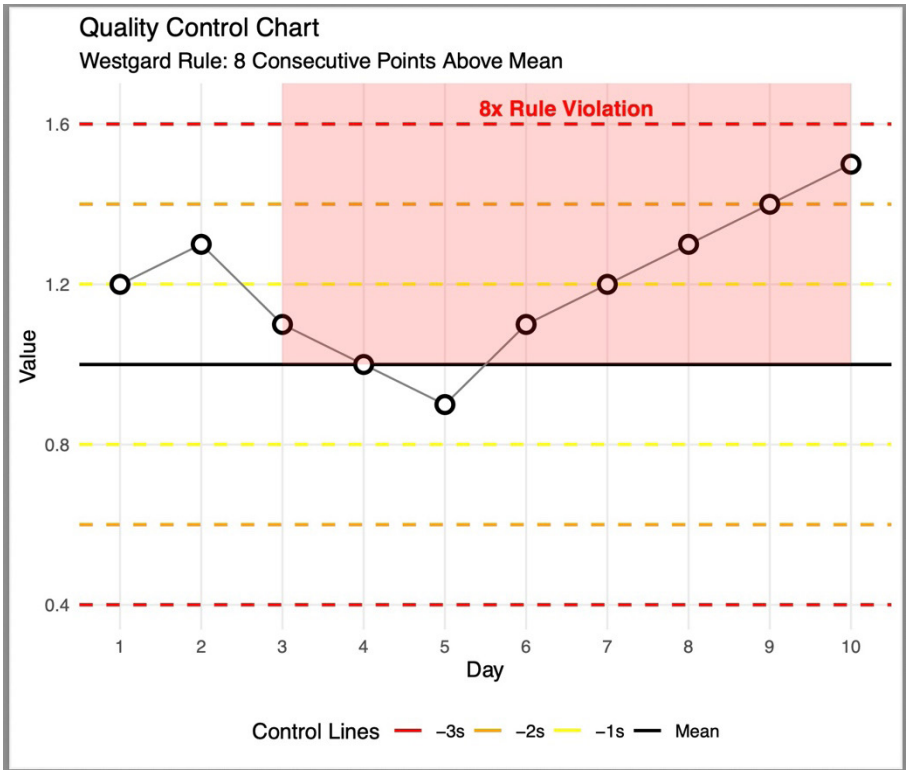

Figure 2: an 8x rule violation means consecutive 8 runs exceed the limits.

However, the limitation imbalance is always a tricky problem in this system. If QC tests a narrow range, such as 12s, the high sensitivity can better detect real problems, but normal variability may frequently trigger alarms. It increases the false positive rate and manual workload. By contrast, high specificity followed by a wider range will avoid false alarms effectively, while certain systematic biases or small shifts may be missed due to low sensitivity. Recently, serial testing and parallel testing have the potential to help optimize the system. Serial testing is used to measure specificity based on the principle that false alarms are minimized by using 12s as a warning rule. Next, confirming any problems by applying more specific rules that have a low probability of false rejection. Parallel testing is used to measure sensitivity, since under the control, the true alarms are maximized by selecting a combination of the rules most sensitive to the detection of random and systematic errors. If any of the rules are violated, the run is rejected.

3. The emergence of real-time data

As time passes, the presence of Patient-Based Real-Time Quality Control achieves a breakthrough by evaluating real-time patient data instead of quality control materials in traditional QC. Quality control materials are stable materials for patient samples and are typically run before patient samples each day to confirm instrument and reagent performance. In blood glucose testing, the lab may use low-level serum QC to simulate hypoglycemic conditions and high-level serum QC to ensure the system can detect hyperglycemia normally.

Simple Moving Average (SMA) and Exponential Weighted Moving Average(EWMA) are two statistical tools for time-series data monitoring used in PBRTQC and are effective for real-time updating. Moving Average is the arithmetic mean of observations within a fixed window length k, used to reduce data fluctuations and highlight trends. SMA is equal weight for each point and can detect chronic changes, such as systematic drift [2]. Compared to more advanced EWMA, SMA has relatively low response speed. EWMA assigns greater weight to more recent data points and assigns a smaller weight to older data points, enabling more sensitive tracking of changes in test data and more suitable for an automated system.

Traditional clinical lab QC methods, such as internal QC, external QA, and PBRTQC, assume patient results are independent and identically distributed, and follow the same overall statistical pattern. Lab test orders are not independent and vary based on several situations. Morning samples may cluster by inpatient status, and certain days of the week might have different patient populations, such as age, sex. This unavoidable situation violates independence, making standard QC methods less reliable.

Here is one real-time example of running glucose tests in a hospital. In scenario 1, all the data are ideally independently distributed:

· The glucose level of Patient A is 95 mg/dL
· The glucose level of Patient B is 110 mg/dL
· The glucose level of Patient C is 87 mg/dL

These values come from random individuals from the same population and are not influenced by when and where the test was done.

In scenario 2, the data are not independently distributed: From 5:00 AM to 7:00 AM, the lab processes samples from ICU patients, who usually have very high glucose due to illness or IV glucose; From 9:00 AM to 12:00 PM, the lab receives samples from outpatients, mostly healthy, with normal glucose levels.

Under real conditions, the data are clustered by patient type and time, and the distribution is skewed. As a result, the QC models that assume all the data are independently distributed may trigger false alarms in the morning due to naturally high glucose from the ICU, since they cannot learn abnormal patterns or adapt to the real condition. Moreover, they also miss true errors because variation from patient mix can mask actual instrument problems.

4. Machine learning in QC

To overcome these challenges, machine learning has great potential to play a role in quality control. By fusing advanced algorithms into the system, it can learn complex patterns from large-scale patient data rather than relying on predefined thresholds. MLQC is proven to have higher accuracy and a faster detection rate for lymphocytes. Data shows that MLQC detected error only after 5 patients, compared to a minimum of 72 patients for the PBRTQC model[3]. The model with fewer patients was affected before bias was detected, which is considered better performance. The MLQC model highlights the effective use of delta data to generalize performance. Several significant mechanisms guide the operation of MLQC: Inputting with the model with lab values, delta data, and patient information; generating output labeled with normal or biased; collecting data for blood counting; preprocessing with an isolation forest (IF) to detect outliers and remove abnormal patterns before feeding into the RF model; and introducing intentional bias during validation and testing. The excellent logistical pattern for the RF model helps categorize the non-linear data very well and automatically captures feature interactions.

The CSLR model is a logistic regression model to estimate the probability of error for each result. It uses stepwise linear regression to predict each analyte using the other 13. CSLR's great innovation is incorporating time of day and day of week to capture daily test variation and introduce simulated biases into training data to evaluate detection power. Also, the system triggers an investigation when the cumulative error score exceeds a threshold. CSLR model can find bias using fewer samples compared to a simple model, which misses 39% of the errors, and has greater precision that detects 98% of all simulated errors [3].

To better understand how machine learning functions within quality control, it is fundamental to consider training methods, the sources of data, and the structure of the model. A robust model that can fit the extensive needs of clinical laboratory and can detect various kinds of errors: interference from poor serum quality, unexplained large result changes, contamination from IV fluids or EDTA, sample degradation, and biologically impossible results. An artificial neural network(ANN) is an eye-catching model that operates with the guidance of artificial intelligence and machine learning. ANN is notable for its robust automation capability. To be specific, a greater proportion of samples are judged by the model as "valid" and saves the time of manual review. In clinical laboratories, test results are reviewed to decide which samples are to be directly released, and those that remain are considered failed. Its accuracy is shown by a high pass rate, 87%-94% of samples directly passing compared with a 50.2%-65.1% in the baseline system, and an AUROC value up to 0.998[3]. AUROC, for binary classification (e.g., valid vs. invalid, healthy vs. diseased), measures the model's ability to correctly distinguish between a randomly chosen positive sample and a randomly chosen negative sample. The closer the AUROC value is to 1, the stronger the distinction between positive and negative. For example, if the AUROC value equals 0.5, the model is like flipping a coin and randomly choosing a label [4].

Another model, known as a Convolutional Neural Network (CNN), has developed an unbelievable self-learning ability attributed to advanced artificial intelligence. CNN can automatically learn patterns or features from raw input using small filters, like sliding windows, and look at the data locally, detect important trends or spikes, and build up a global understanding. Four major steps in the CNN system lead the model:

1. Local receptive fields: use convolutional filters that slide like a moving window across a time series. Each filter focuses on a local pattern of several consecutive time points, detecting a rapid rise or periodic fluctuation. If a test value (like blood glucose) increases from 100 → 105 → 200, this sudden change within a short time span can be captured by the CNN.
2. Weight sharing: CNN filters are shared across the entire time series means they look for the same type of pattern

ISSN 2959-409X

throughout all periods. Whether the spike appears in the morning or afternoon, CNN can still detect it. It eliminates the need to manually inform the model about phenomena like summer variation or morning spikes.

3. Multi-layer architecture: each layer learns more abstract and complex features. The first layer detects small fluctuations, and the second layer might capture long-term increasing trends. The third layer identifies anomalous patterns. Traditional statistical models require explicitly telling the model "consider seasonality" or adding a time term. In PBRTQC, CNN can detect time-dependent anomalies in lab data to find hidden patterns that indicate an out-of-control event. It saves the effort of manually labeling the special seasonality or period.

4. No feature engineering needed: traditional statistical models require explicitly telling the model " consider seasonality" or adding a time term. Staff only need to input raw time-series data, and the model learn what matters and what patterns exist.

To reduce the influence of extreme values on overall data calculation, AI-driven Winsorization achieves this goal by replacing extreme values with less extreme ones [5]. It is an effective way to process extreme patient test results, like abnormal results caused by rare diseases, preventing these values from skewing the moving average or moving median, which could otherwise lead to false alarms in the system. In PBRTQC, it is used to automatically find the optimal combination of parameters and optimize according to a cost function, such as a lower false alarm rate and smaller Nped (number of patients affected before error detection). However, AI-driven Winsorization has not been widely adopted among clinical labs due to several challenges. Its advanced technology raises high requirements for staff familiarity and complex connectivity issues. Also, there is a large amount of training data needed for model optimization for smaller laboratories.

5. Challenges

Machine learning has great potential to promote the development of PBRTQC, however, the limitations it is facing now are worth deeper discussion and research.

If Biases exist when data sets are created, they will be embedded into the training process and generate a biased model. The biases can be missing data or imbalanced data. Missing data will cause a shifted baseline and reduced sensitivity to the model. For example, A PBRTQC system that monitors creatinine levels to detect analytical bias. It gains data from two populations: inpatients with relatively higher creatinine due to acute kidney injury and outpatients, usually accompanied by lower creatinine values. Due to connectivity issues, output results are not being transmitted to the PBRTQC database for several weeks, and the model is trained only with inpatient results. As a result, the system learns that higher creatinine values are normal and small upward drifts in analyzer performance are masked because they fall within this artificially high baseline.

Imbalanced data can be specified as the scarcity of abnormal samples in the real situation, so that many algorithms train exclusively on negative samples, the results from healthy patients, or accurate test outputs. Any samples that deviate from these learned characteristics are judged as a positive, such as an error or unexpected behavior.

Subgroup-specific model is also an important factor that should be taken into consideration for the system's comprehensive performance [6]. Since this model excludes samples in another subgroup, leads actual performance was overestimated. This phenomenon also exists in multi-instrument laboratories, indicating that using independent models may not accurately reflect overall testing performance. It can be explained as two groups, one is hospitalized patients, and the other are ambulatory patients. Each group has separate PBRTQC rules or thresholds. The underlying problem is the performance evaluation—the average number of patients affected before error detection(ANped)— was only calculated within the same group. It means concurrent samples are excluded from the other subgroup that was processed in parallel in the lab. But in real lab operations, both subgroups' samples are processed together, and errors or drifts often affect all patient samples across subgroups. By testing only within one subgroup, the model may detect errors faster, but only because it's not looking at the full picture. The model might seem to detect errors with high sensitivity, but that performance doesn't reflect what would happen in a real, mixed dataset. Inter-subgroup or instrument-wide errors may go unnoticed if models don't account for the whole system.

One persistent and often ignored challenge impeding the optimization of the AI-driven PBRTQC is the reliance on simulated datasets for model training. It uses artificial errors that abruptly change test values to create an "out-of-control" example. In real time, transitions from in-control to out-of-control are often gradual and vary in form. As a result, models trained on these artificial patterns may fail to detect real-world issues that change slowly over time.

6. Future projection

New studies point out that DeepSeek plays an indispensable role in promoting the development of PBRTQC by using learning and natural language processing technologies to quickly analyze, identify, and extract key infor-

mation from reports, enabling automated report review. DeepSeek uses streaming data plus predictive models to detect analyzer drift or failure before errors occur. Traditional QC methods rely on fixed thresholds, often with high false alarm rates.

Dynamic quality rule optimization can also be achieved by DeepSeek since it can adjust QC rules weekly basis based on new data, compared to a 6-12months update frequency for traditional QC rules. THE AI QC lock mechanism is also a competitive advantage that automatically flags and locks the abnormal results using AI algorithms, reducing manual misinterpretation risk. Traditional PBRTQC relies heavily on human review and simple SPC algorithms. Moreover, by linking patient symptoms, history, and clinical guidelines, DeepSeek recommends an appropriate test combination and detects when results don't fit the expected pattern[7].

A hybrid approach with a rules-based system can be an innovative way to improve the precision of the AI-driven PBRTQC. The hybrid approach combines ML with existing rules-based algorithms to reach ideal performance. To ensure both immediate accuracy and long-term stability, it is essential to integrate human oversight with continuous performance monitoring from the very start of deployment. Human-in-the-loop review means all ML-driven decisions should be reviewed by a human before final action, especially during early integration. Monitor and feedback loops means a feedback system where model errors are tracked and used to restrain the model is applied. Here are two hybrid examples: Mindray's NN-PBRTQC adopts an ANN mapping model that integrates patient-specific parameters, including sex, age, clinical department, and disease diagnosis, to predict test values. ANN learn hidden relationships between the influencing factors [3]. It can also learn automatically, no need for data expertise, alarm filter to reduce the false alarm rate. mNL-PBRTQC by Zhou uses a nonlinear classification and regression tree regression model, incorporating the test date, test time, instrument brand, and hospital level [3].

7. Conclusion

The integration of AI-driven methods into Patient-Based Real-Time Quality Control (PBRTQC) represents a significant advancement in clinical laboratory error detection, offering improved sensitivity, adaptability, and operational efficiency compared to traditional QC approaches. From the evolution of Westgard rules to machine learning models such as CSLR, ANN, and CNN, these systems demonstrate the ability to process complex, high-dimensional patient data and identify subtle performance drifts more rapidly. However, the persistent challenges of biased training data, subgroup limitations, unrealistic simulated datasets, and data imbalance highlight the need for continued refinement. Emerging tools like DeepSeek and hybrid AI–rules-based frameworks provide promising pathways for dynamic optimization, reduced false alarms, and enhanced real-world applicability. Ultimately, addressing these limitations while leveraging AI's analytical strengths will be essential for achieving reliable, scalable, and clinically integrated PBRTQC systems that improve patient safety and laboratory performance.

References

[1] Westgard, J. O. (2016). *Basic QC Practices: Training in Statistical Quality Control for Medical Laboratories* (4th ed.). Madison, WI: Westgard QC. ISBN 978-1-886958-30-2.

[2] Fernando, J. (n.d.). Moving average (MA): Purpose, uses, formula, and examples. Investopedia. https://www.investopedia.com/terms/m/movingaverage.asp

[3] Lorde, N., Mahapatra, S., & Kalaria, T. (2024). Machine Learning for Patient-Based Real-Time Quality Control (PBRTQC), Analytical and Preanalytical Error Detection in Clinical Laboratory. *Diagnostics*, *14*(16), 1808. https://doi.org/10.3390/diagnostics14161808

[4] Mastering AUC-Roc: Essential Model Evaluation Techniques. Galileo. (n.d.). https://galileo.ai/blog/auc-roc-model-evalulation

[5] Author links open overlay panelRand R. Wilcox. (n.d.). Summarizing data. Applying Contemporary Statistical Techniques. https://www.sciencedirect.com/science/article/abs/pii/B9780127515410500249

[6] Duan X, Zhang M, Liu Y, Zheng W, Lim CY, Kim S, Loh TP, Guo W, Zhou R, Badrick T; Patient-Based Real-Time Quality Control Working Group of the Asia Pacific Federation of Clinical Biochemistry and Laboratory Medicine. Next-Generation Patient-Based Real-Time Quality Control Models. Ann Lab Med. 2024 Sep 1;44(5):385-391. doi: 10.3343/alm 2024.0053. Epub 2024 Jun 5. PMID: 38835211; PMCID: PMC11169771.

[7] Senyu Senyu Medicine result Ai-PBRTQC - Wisdom test. (2025, February 16). https://www.sxmedical.net/gongsixinw/106.html