IFCC Paper

Tze Ping Loh*, Andreas Bietenbeck, Mark A. Cervinski, Huub H. van Rossum, Alex Katayev and Tony Badrick, on behalf of the International Federation of Clinical Chemistry and Laboratory Medicine Committee on Analytical Quality

Recommendation for performance verification of patient-based real-time quality control

<https://doi.org/10.1515/cclm-2019-1024>

Received October 4, 2019; accepted January 13, 2020

Abstract: Patient-based real-time quality control (PBRTQC) is a laboratory tool for monitoring the performance of the testing process. It includes well-established procedures like Bull's algorithm, average of nomals, moving median, moving average (MA) and exponentially (weighted) MAs. Following the setup and optimization processes, a key step prior to the routine implementation of PBRTQC is the verification and documentation of the performance of the PBRTQC as part of the laboratory quality system. This verification process should provide a realistic representation of the performance of the PBRTQC in the environment it is being implemented in, to allow proper risk assessment by laboratory practitioners. This document focuses on the recommendation on performance verification of PBRTQC prior to implementation.

Keywords: evaluation; moving average; moving median; patient-based quality control; patient-based real-time quality control; verification.

Background

Patient-based real-time quality control (PBRTQC) is a laboratory tool for monitoring the performance of the testing process. It includes well-established procedures like Bull's algorithm, average of nomals, moving median, moving average (MA) and exponentially (weighted) MAs [1–7]. More recently, novel techniques such as the moving standard deviation, moving delta, moving sum of outliers and moving percentiles have been described [8–10]. These techniques have gained increasing attention owing to maturing statistical methodology, improved information technology capabilities and increasing awareness of the limitations of internal quality control systems [9, 11–16]. Indeed, Bull's algorithm (a form of average of normals) is routinely used in clinical hematology laboratories.

Recent successful implementations and proof of value of PBRTQC in complex laboratories have further given confidence in this technique [17–19]. The cost savings and potential ability to withhold results until the verification of performance of the testing system are further advantages that fit well in the austere and risk-aware climate in laboratory medicine practice.

The International Federation of Clinical Chemistry and Laboratory Medicine PBRTQC Working Group has recently produced separate documents that provide guidance on the informatics considerations [20] and implementation of PBRTQC [21]. These documents serve to facilitate the adoption of PBRTQC in routine laboratories, and readers are strongly encouraged to read them to familiarize themselves with key concepts before this document. A key step prior to the routine implementation of PBRTQC is the verification and documentation of the performance of the PBRTQC as part of the laboratory quality system. This verification process should provide a realistic representation of the performance of the PBRTQC in the environment it is being implemented in, to allow proper risk assessment by the laboratory practitioners. This document focuses on the recommendation on performance verification of PBRTQC prior to implementation.

***Corresponding author: Tze Ping Loh**, Department of Laboratory Medicine, National University Hospital, 5 Lower Kent Ridge Road, Singapore 119074, Singapore, Phone: (+65) 67724345; Fax: (+65) 67771613, E-mail: tploh@hotmail.com

Andreas Bietenbeck: Institut für Klinische Chemie und Pathobiochemie Klinikum, Munich, Germany.
<https://orcid.org/0000-0002-1228-0770>

Mark A. Cervinski: Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA; and Geisel School of Medicine at Dartmouth, Hanover, NH, USA

Huub H. van Rossum: The Netherlands Cancer Institute, Amsterdam, the Netherlands; and Huvaros, Amsterdam, the Netherlands

Alex Katayev: Laboratory Corporation of America Holdings, Elon, NC, USA

Tony Badrick: RCPA Quality Assurance Programs, St Leonards, Sydney, Australia

Data extraction

It is important that the laboratory plans the data extraction such that separate datasets are allocated for data familiarization, parameter setting and optimization purposes (training set), as well as for the performance verification (verification set). Generally, the data are extracted over a continuous period and partitioned. The partitioning ratio of the training to verification set can range from 50:50 to 80:20, depending on data density. It is preferable to allocate the majority of the extracted data for parameter setting and optimization, and leave the remaining data for performance verification. The dataset can be portioned by simple partitioning, or using more sophisticated randomization techniques. The key consideration is to ensure that the verification set is large enough to yield meaningful results and is representative of the dataset as a whole.

Representative historical data are extracted as training dataset to learn the pattern of the measurand in the population served by the laboratory and setting the PBRTQC parameters. The historical data should cover a sufficient duration of data collection to capture all the potential variability of the measurand, and may be dependent on the clinical setting and laboratory method. Important variability to consider includes patient population variability, reagent lot and calibrator lot changes (at least two different lots). Generally, at least 6 months of data should be collected (to cover sufficient weekends), although data collected over a year or more will be more likely to comprehensively capture all variability, including circannual changes. At the same time, for highly specialized practices (e.g. a cancer institution), care must also be exercised not to collect data from periods before there was a change in clinical practice (e.g. a change in cancer treatment protocol) that may significantly affect the distribution of some measurands. Inclusion of such historical data may unnecessarily increase the variability in the result distribution, confounding the setup parameters and verification interpretation. Any known periods with potentially erroneous results due to laboratory error should be excluded.

Setup of patient-based real-time quality control

Guidance on the setup of PBRTQC has been described in detail previously [21, 22]. Briefly, the laboratory user should familiarize themselves with the relevant physiological, pathological and pre-analytical factors that can influence the measurand of interest and with the

variations in ordering behavior that change the distribution of results. The optimal block size, statistical model, specific inclusion/exclusion criteria, truncation limits and data transformation are applied, as necessary. The mean of the population or population result blocks and standard deviation are calculated and control limits are chosen based on selected quality goals. The PBRTQC parameters can be selected based on published examples, power function graphs or simulation software (e.g. <https://www.huvaros.com>) [17, 23].

Performance verification

After the basic setup of the PBRTQC parameters, they need to be subjected to performance verification that reflects the local laboratory setting. This is to ensure proper settings are implemented before going live and to avoid disruption to routine clinical operations. It can also serve as a familiarization exercise and a check on the proper functioning of the software. The findings of the performance verification can be used to further refine the PBRTQC parameters and optimize the performance that best balances the risk profile and operational requirements of the laboratory. It should be noted that while minor adjustments to the PBRTQC may be considered acceptable, significant changes in the parameter warrant a repeat verification with a separate set of historical data. The final selected PBRTQC parameters should be documented along with the performance characteristics obtained during the verification process.

Types of performance verification

There are several measures of performance for PBRTQC, and they have been reviewed previously [24]. They include power function analysis, total error allowable (TEa) detection probability, average or median number of patient results affected before error detection (ANPed or MNPed), sensitivity/specificity of error detection, bias detection curves and number of patient results required to detect error with a given probability. The performance verification of PBRTQC can be undertaken by *in silico* analysis. This can be performed without incurring a high cost beyond the manpower and computational power required.

This document provides detailed guidance on the ANPed method due to its relative simplicity in concept and application, which can be accomplished using common statistical software such as Microsoft Excel.

Additionally, verification methods that are based on more complex approaches such as bias detection curves and MA validation charts are also discussed [23]. These approaches provide a patient risk-based assessment of the PBRTQC setup and are based on realistic error detection simulations.

Establishing false-positive rate

The curated verification laboratory data are first arranged in chronological order according to the time of reporting. Following this, the selected PBRTQC parameters are applied to the verification dataset. This should yield a baseline PBRTQC model, an example of which is graphically represented in Figure 1. The number of instances the MA exceeds the lower or upper control limits (flags) is recorded. In the example shown in Figure 1, there were seven flags noted. This represents a simplistic false-positive flag rate of 0.7% (7/1000). In practice, the number of flags can be easily calculated by filtering and counting the number of results lying outside of the control limits. Depending on the measurand, the laboratory should determine the desired number of false alarm (false-positive rate) per period (e.g. days or weeks) that the laboratory deems as manageable [23]. Others have used an approach in which, by design, automatically the control limits are based on allowing no false alarms in the training set, which can be adjusted manually [23]. The verification process should ensure that the desired false-positive rate is not exceeded.

Figure 1: A patient-based real-time quality control chart of moving average of 1000 consecutive sodium measurements with a block size of 20.

The arrows denote the false-positive flags in the absence of laboratory errors. UCL, upper control limit; LCL, lower control limit.

In reality, when a MA alarm is triggered, it will be investigated and the affected results will be discarded from future analysis to avoid triggering false alarms after resolution of the underlying issue. To better define the false-positive rate operationally, it can be expressed as the number of days (or shifts) with flags over the number of days (or shifts) without flags. This requires more nuanced analysis that takes into consideration the operational conditions of the laboratory and the data handling during analysis.

Establishing the ANPed

Systematic error (bias) can be introduced after the baseline PBRTQC model is set up as mentioned earlier. In this example, a systematic error of 3 mmol/L, which represents the analytical performance specification (or “total allowable error”) for sodium recommended by The Royal College of Pathologists of Australasia Quality Assurance Programs, is selected. Other magnitudes (e.g. in fractions or multiples of analytical performance specification) of systematic error can be examined depending on the quality goals of the laboratory.

To simulate a positive shift, the magnitude of error under examination is added to the original results of the model, following which the PBRTQC model is reapplied and the moving statistics are recalculated. It is emphasized that the systematic error should be introduced before any data truncation and/or transformation is applied. It is desirable to examine the systematic error in both the positive and negative directions, as either may happen in practice. Moreover, the PBRTQC error detection performance can differ significantly between positive and negative bias.

The selected systematic error can be introduced at different time points in the dataset. This can be achieved by introducing (i.e. adding or subtracting) the systematic error after n th results in the curated verification dataset and reapplying the PBRTQC model, as shown graphically in Figure 2A. The systematic error can be introduced and sustained for a variable number of block sizes to simulate variable duration of sustained error (e.g. from one block size to more than 10 block sizes). For the current example, 3 mmol/L is added to 200 consecutive results (10 times the block size of 20), after the result that corresponded with the 20th data point in the MA.

The interval between the point of introduction of the systematic error (the 21st MA data point) and the first MA data point that breaches the control limit (the 31st MA

Figure 2: An example of the number of patient results affected before error detection (NPed). (A) Introduction of a systematic error (3 mmol/L) into the base patient-based real-time quality control model after the 20th result and sustained for 200 subsequent results. (B) The process is repeated where the systematic error is introduced after the 40th result and sustained for 200 subsequent results. Panel (C) shows the introduction of systematic error in the negative direction. The number of patient results affected before error detection was 10, 11 and 7 for panels A, B and C, respectively. UCL, upper control limit; LCL, lower control limit.

data point) is the number of patient results affected before error detection (NPed=10, see Figure 2A). The systematic error can be introduced at multiple subsequent time points at any interval in the verification dataset and the NPed can be derived (Figure 2B). The same process should be repeated for systematic error in the negative direction (Figure 2C). The aforementioned assessment can be coded as a macro or other simulation set up to automate the process.

The systematic error can be introduced repeatedly at any interval. However, it is generally desirable to have intervals evenly distributed across the entire dataset to ensure comprehensive coverage of the variability in the dataset. It is also desirable to perform sufficient number of repeat introductions of the systematic error for a given dataset to ensure robust statistical estimation. The average of the NPed derived from the repeat introduction of systematic error is the ANPed.

Assessment of ANPed

The desirable ANPed is specific to the measurand and depends on the clinical risk of harm of the measurand and the risk profile of the laboratory. Smaller ANPed indicates a lower risk of erroneous patient result reporting prior to error detection when “report from the back” (where the first sample analyzed is not released until the remainder of the MA block is analyzed and found not to exceed the control limits) strategy is not employed [18]. However, ANPed generally has an inverse relationship with false-positive flag rate. Care must be taken to balance the benefit of a low ANPed (earlier error detection) and the risk of a high false-positive flag rate, where a significant laboratory resource may be expended to address the false alarm. A high false alarm rate may also compromise the confidence in the PBRTQC system, leading to alarm fatigue. Worse, a false alarm may unnecessarily delay the reporting of laboratory results that may provide crucial clinical information for patient care.

In general, it is desirable to have an ANPed that is smaller than the number of patient samples analyzed between the internal quality control testing in use in the laboratory or smaller than the given results block size. This improves the error detection (and harm/risk reduction) capability of the laboratory over existing internal quality control practice. Alternatively, it may be desirable to have the ANPed smaller than the time it takes to correct any systematic error and prevent release of results which could lead to patient harm. An actual verification exercise performed in the laboratory of one of the authors is shown in Figure 3.

A. Positive bias of 30% introduced starting from sample 301. Sensitivity A = 100%; Sensitivity B = 100% (NPed less than block size)

B. Negative bias of 30% introduced starting from sample 301. Sensitivity A = 100%; Sensitivity B = 100% (NPed less than block size)

Figure 3: Example of a verification study in a clinical laboratory.

A series of production results that were run on selected instrument channels were extracted from the laboratory database. Next, the artificial bias was introduced (positive [panel A] in one experiment and negative [panel B] in the other experiment) via slope correction and both, unbiased and biased results were run offline using a simulation program. The error detection sensitivity was assessed in two dimensions: ability of the erroneous results to trigger moving median rules (flag error thresholds – Sensitivity A) and reliability of the “release from the back” (where the first sample analyzed is not released until the remainder of the MA block is analyzed and found not to exceed the control limits); the last result within a block is an algorithm not to allow any erroneous results to be released (Sensitivity B). Blue circles = calculated blocks of patient results, red lines = error limits, green line = mean of blocks (cumulative from all labs), measurand = free thyroxine, algorithm for block calculation = mean of $\ln(\text{result} \times 1000)$.

Figure 4: PBRTQC optimization and validation using hemoglobin as an example.

Upper; PBRTQC optimization using bias detection curves. Here the performance of five PBRTQC protocols calculating the mean of the last 5, 10, 25, 50 and 99 are presented. The lines present the median number of test results needed for PBRTQC error detection (Y-axis) when a range of systematic errors (X-axis) was studied, using the corresponding PBRTQC settings. This way the error detection performance of various PBRTQC procedures can be compared and the most “optimal” error detection profile can be selected. This process can be repeated several times to study all variables of interest (algorithm, truncation limits, etc.). Most often, optimization is based on balancing rapid detection of large error versus enabling detection of smaller error or reliable detection of TEa. Lower; MA validation chart. After selection of the optimal PBRTQC (here the PBRTQC protocol using a mean calculation of last 25 results), the performance verification is presented in a MA Validation chart. In the MA Validation chart also, the uncertainty systematic error detection by PBRTQC is presented. In the MA Validation chart, the bars represent the median number of results needed for systematic error detection and error bars present the 95% interval of the simulation results. Therefore, the upper error bars represent the number of patient results needed to enable detection of the systematic error in 97.5% of the performed simulations. Data were obtained using the MA Generator (www.huvaros.com. Accessed 30 September 2019).

Bias detection curves and MA Validation charts

An alternative approach to optimize and verify PBRTQC settings uses a similar simulation design but calculates the median number of patient results affected [23]. These simulations are performed for different systematic errors and PBRTQC calculation algorithms, and the obtained results are presented in so-called bias detection curve plots. These plots can be used to compare the performance of the different PBRTQC protocols and allow selection of the PBRTQC protocol that is considered most optimal for the lab (Figure 4). The next step is to obtain a more thorough verification of the selected optimal PBRTQC protocol by obtaining a MA validation chart. Here, not only the median but also the 95%, 99% or min/max number of patient results affected, as obtained from the performed simulations, are presented (Figure 4). Therefore, this

comprises a more thorough insight into the PBRTQC error detection performance and perhaps more importantly the uncertainty thereof. The use of these graphs is supported by the online MA Generator application available for laboratories (www.huvaros.com. Accessed 30 September 2019). The appropriateness of this approach has been demonstrated in a prospective study that demonstrated a manageable number of alarms when applied for 24 chemistry tests, each run on two random-access chemistry platforms. Furthermore, PBRTQC allowed detection of relevant errors [23].

Recently, an implementation approach that incorporated the use of the MA Generator to allow PBRTQC implementation was presented. This approach included, amongst others, the next steps for successful PBRTQC implementation, selection of assays for which to consider PBRTQC, application of the MA Generator to obtain PBRTQC settings, a verification phase to confirm PBRTQC

Table 1: Suggested items to document in the patient-based real-time quality control (PBRTQC) performance verification.

Items	Description
PBRTQC parameters	
Background about the measurand	Brief summary of relevant physiological, pathological and pre-analytical factors that can influence the measurand of interest and with the variations in ordering behavior that change the distribution of results
Data source	Source of the data used in the development of the PBRTQC model, e.g. from the laboratory information system
Date of extraction	Date when the data was extracted
Duration of data for training set	Period covered by the extracted data for PBRTQC parameter setup and optimization
Duration of data for verification set	Period covered by the extracted data for PBRTQC performance verification
Inclusion criteria	Any inclusion criteria used by the PBRTQC model
Exclusion criteria	Any exclusion criteria used by the PBRTQC model
Truncation limits	Truncation limits used to exclude outliers or reduce variability in the data, provide source as well
Data transformation	Statistical transformation applied to normalize the data (i.e. to approximate a Gaussian distribution)
Moving statistics used	Moving statistics used, e.g. moving average, moving median, etc.
Population target value	Mean value of the population or population result blocks
Control limit	Generally taken as multiples of the standard deviation of the mean of the population or population result blocks
Last revision date	Documentation of the revision date
Performance verification	
Data source	Source of the data used in the verification of the PBRTQC model, e.g. from the laboratory information system
Date of extraction	Date when the data was extracted
Duration of data	Period covered by the extracted data
Systematic error selected	Magnitude of systematic error selected for examination, provide source or rationale if possible (e.g. analytical performance specification of a certain quality assurance program or by biological variation)
Statistical software used	Statistical software used for the verification exercise, provide version number
Person performing the verification	Name of staff who performed the verification
Verification parameter selected	Verification method selected, for example average number of patient results affected before error detection
Achieved verification parameter	Outcome of the performance verification expressed in a suitable statistical parameter, e.g. false-positive rate, ANPed, etc.

appropriateness and the design of laboratory protocols to enable proper acknowledgment of PBRTQC alarming [24].

Other verification methods

Many alternate strategies have been described for verification of the performance of PBRTQC. They may be explored according to the comfort and familiarity of the laboratory with the underlying statistical methodologies that have been described elsewhere [22]. In general, they are all valid methods of performance verification, although methods that provide information about patient risk may be preferred. A specialized simulation software may provide the most convenient way of performing this process (e.g. <https://www.huvaros.com>). Another useful method of verifying the performance of PBRTQC is to run the algorithm in the production environment without activating the alarms. This will help give the laboratory a realistic picture of the number of alarms that will be generated, and decide whether it is manageable before going live.

Documentation of performance verification

Upon finalization of the PBRTQC parameters and completion of the verification process, both sets of parameters should be properly documented. This will help ensure traceability of the rationale, process of adopting a particular PBRTQC setup, and regulatory compliance. Suggested items to document are shown in Table 1.

Conclusions

PBRTQC is an exciting new tool in the repertoire of quality tools in the laboratory. This document serves as a practical guide to performance verification of PBRTQC that will provide confidence in the method as well as proper documentation for regulatory purpose. It is hoped that it will reduce the perceived barrier to adopting this technique in routine laboratory practice.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Employment or leadership: Huub H. van Rossum has a financial interest in Huvaros.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

- Hoffmann RG, Waid RE. The "Average of normals" method of quality control. *Am J Clin Pathol* 1965;43:134–41.
- Bull BS, Elashoff RM, Heilbron DC, Couperus J. A study of various estimators for the derivation of quality control procedures from patient erythrocyte indices. *Am J Clin Pathol* 1974;61:473–81.
- Cembrowski GS, Chandler EP, Westgard JO. Assessment of "Average of Normals" quality control procedures and guidelines for implementation. *Am J Clin Pathol* 1984;81:492–9.
- Smith FA, Kroft SH. Exponentially adjusted moving mean procedure for quality control. An optimized patient sample control procedure. *Am J Clin Pathol* 1996;105:44–51.
- Neubauer AS. The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation. *Clin Chem* 1997;43:594–601.
- Linnet K. The exponentially weighted moving average (EWMA) rule compared with traditionally used quality control rules. *Clin Chem Lab Med* 2006;44:396–9.
- Bietenbeck A, Thaler MA, Luppa PB, Klawonn F. Stronger together: aggregated Z-values of traditional quality control measurements and patient medians improve detection of biases. *Clin Chem* 2017;63:1377–87.
- Jones GR. Average of delta: a new quality control tool for clinical laboratories. *Ann Clin Biochem* 2016;53:133–40.
- Liu J, Tan CH, Badrick T, Loh TP. Moving sum of number of positive patient result as a quality control tool. *Clin Chem Lab Med* 2017;55:1709–14.
- Liu J, Tan CH, Badrick T, Loh TP. Moving standard deviation and moving sum of outliers as quality tools for monitoring analytical precision. *Clin Biochem* 2018;52:112–6.
- Howanitz PJ, Tetrault GA, Steindel SJ. Clinical laboratory quality control: a costly process now out of control. *Clin Chim Acta* 1997;260:163–74.
- Miller WG, Ere A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. *Clin Chem* 2011;57:76–83.
- Algeciras-Schimnich A, Bruns DE, Boyd JC, Bryant SC, La Fortune KA, Grebe SK. Failure of current laboratory protocols to detect lot-to-lot reagent differences: findings and possible solutions. *Clin Chem* 2013;59:1187–94.
- Loh TP, Lee LC, Sethi SK, Deepak DS. Clinical consequences of erroneous laboratory results that went unnoticed for 10 days. *J Clin Pathol* 2013;66:260–1.
- Thaler MA, Iakubov R, Bietenbeck A, Luppa PB. Clinically relevant lot-to-lot reagent difference in a commercial immunoturbidimetric assay for glycated hemoglobin A1c. *Clin Biochem* 2015;48:1167–70.

16. Koerbin G, Liu J, Eigenstetter A, Tan CH, Badrick T, Loh TP. Missed detection of significant positive and negative shifts in gentamicin assay: implications for routine laboratory quality practices. *Biochem Med (Zagreb)* 2018;28:010705.
17. Ng D, Polito FA, Cervinski MA. Optimization of a moving averages program using a simulated annealing algorithm: the goal is to monitor the process not the patients. *Clin Chem* 2016;62:1361–71.
18. Fleming JK, Katayev A. Changing the paradigm of laboratory quality control through implementation of real-time test results monitoring: for patients by patients. *Clin Biochem* 2015;48:508–13.
19. van Rossum HH, Kemperman H. Moving average for continuous quality control: time to move to implementation in daily practice? *Clin Chem* 2017;63:1041–3.
20. Loh TP, Cervinski MA, Katayev A, Bietenbeck A, van Rossum H, Badrick T, International Federation of Clinical Chemistry and Laboratory Medicine Committee on Analytical Quality. Recommendations for laboratory informatics specifications needed for the application of patient-based real time quality control. *Clin Chim Acta* 2019;495:625–9.
21. Badrick T, Bietenbeck A, Cervinski MA, Katayev A, van Rossum HH, Loh TP, International Federation of Clinical Chemistry, and Laboratory Medicine Committee on Analytical Quality. Patient-based real-time quality control: review and recommendations. *Clin Chem* 2019;65:962–71.
22. van Rossum HH. Moving average quality control: principles, practical application and future perspectives. *Clin Chem Lab Med* 2019;57:773–82.
23. van Rossum HH, Kemperman H. Optimization and validation of moving average quality control procedures using bias detection curves and moving average validation charts. *Clin Chem Lab Med* 2017;55:218–24.
24. van Rossum HH, van den Broek D. Design and implementation of quality control plans that integrate moving average and internal quality control: incorporating the best of both worlds. *Clin Chem Lab Med* 2019;57:1329–38.