

Yu-fang Liang, Andrea Padoan, Zhe Wang, Chao Chen, Qing-tao Wang\*, Mario Plebani\* and Rui Zhou\*

# Machine learning-based nonlinear regression-adjusted real-time quality control modeling: a multi-center study

<https://doi.org/10.1515/cclm-2023-0964>

Received August 31, 2023; accepted October 25, 2023;

published online November 21, 2023

## Abstract

**Objectives:** Patient-based real-time quality control (PBRTQC), a laboratory tool for monitoring the performance of the testing process, has gained increasing attention in recent years. It has been questioned for its generalizability among analytes, instruments, laboratories, and hospitals in real-world settings. Our purpose was to build a machine learning, nonlinear regression-adjusted, patient-based real-time quality control (mNL-PBRTQC) with wide application.

**Methods:** Using computer simulation, artificial biases were added to patient population data of 10 measurands. An mNL-PBRTQC was created using eight hospital laboratory databases as a training set and validated by three other hospitals' independent patient datasets. Three different Patient-based models were compared on these datasets, the IFCC PBRTQC model, linear regression-adjusted real-time quality control (L-RARTQC), and the mNL-PBRTQC model.

**Results:** Our study showed that in the three independent test data sets, mNL-PBRTQC outperformed the IFCC PBRTQC and L-RARTQC for all measurands and all biases. Using

platelets as an example, it was found that for 20 % bias, both positive and negative, the uncertainty of error detection for mNL-PBRTQC was smallest at the median and maximum values.

**Conclusions:** mNL-PBRTQC is a robust machine learning framework, allowing accurate error detection, especially for analytes that demonstrate instability and for detecting small biases.

**Keywords:** patient-based real-time quality control; machine learning; nonlinear regression; residual

## Introduction

Since the work of Bull [1] based on the “average of normals” concept described by Hoffman in 1965 (ref), the idea of patient-based real-time quality control (PBRTQC) has been routinely used as a supportive quality assurance tool in hematology. However, implementing PBRTQC into clinical chemistry and immunoassay has been slow. Now, PBRTQC is gaining increasing attention thanks to sophisticated statistical methodologies, improved information technology capabilities and increasing awareness of the limitations of traditional quality control programs (TQC) [2–8]. PBRTQC involves using statistical manipulations of patient results produced from routine clinical analysis [9, 10]. PBRTQC algorithms include already established procedures, moving median (MM), moving average (MA) and exponentially weighted MA (EWMA) [1, 11–16]. Later, the moving standard deviation, moving delta, moving sum of outliers, moving percentiles and Harrell–Davis 50 percentile estimator (HD50) have been further described [2, 3, 10, 17]. Recently, a novel approach for PBRTQC based on machine learning depending algorithms has been reported [18].

Compared with conventional QC, PBRTQC reduces the cost associated with commercial QC materials because it uses patient results already generated; avoids the problems due to the poor commutability of commercial QC materials, increases the sensitivity (true error detection) and specificity (false rejection) [4, 19]; it overcomes an interval-based QC process control and allows for a continuous operational

---

Yu-fang Liang, Andrea Padoan and Zhe Wang contributed equally to this work and should be considered first authors.

---

**\*Corresponding authors: Qing-tao Wang and Rui Zhou,** Department of Laboratory Medicine, Beijing Chao-yang Hospital, Capital Medical University, Beijing, P.R. China, and Beijing Center for Clinical Laboratories, No. 8 Gongti South Road, Chaoyang District, Beijing, 100020, P.R. China, E-mail: wqt36@163.com (Q.-t. Wang), zr-molly@163.com (R. Zhou); and **Mario Plebani,** Department of Medicine-DIMED, University of Padova, Padova, Italy, Phone: +39049663240, Fax: +39049663240, E-mail: mario.plebani@unipd.it. <https://orcid.org/0000-0002-0270-1711> (M. Plebani)

**Yu-fang Liang,** Department of Laboratory Medicine, Beijing Chao-yang Hospital, Capital Medical University, Beijing, P.R. China

**Andrea Padoan,** Laboratory Medicine Unit, University-Hospital of Padova, Padova, Italy. <https://orcid.org/0000-0003-1284-7885>

**Zhe Wang and Chao Chen,** Beijing Jinfeng Yitong Technology Co., Ltd, Beijing, P.R. China

monitoring mode It is gradually becoming an essential tool in the laboratory quality control repertoire. Recently, many efforts have been suggested to optimize the various parameters of PBRTQC, its application in different laboratory settings, and the early error detection [2, 3, 10, 17–22]. A recent study on RARTQC put forward the optimization method by changing the way of data feature selection [23]. Several reviews and recommendations have been published on PBRTQC to offer guidelines to laboratory practitioners [9, 10, 24, 25].

The purpose was to design a novel machine learning nonlinear regression-adjusted patient-based real-time quality control model (mNL-PBRTQC) mainly by optimizing QC decision algorithm to improving its performance and validate its effect in real world as well.

## Materials and methods

### Data collection

The project was organized by Beijing Center for Clinical Laboratories and Shandong Provincial Center for Clinical Laboratories, Hebei Provincial Center for Clinical Laboratories and Tianjin Center for Clinical Laboratories in China. According to China's Hospital Classification, participating hospital laboratories include secondary- and tertiary-level general and special hospital laboratories to ensure data representativeness.

When a laboratory declared its intention to join, it was provided with the Information Technology (IT) requirements for sending data. We verified an error-free transmission into our central database. Subsequently, data transfer occurred automatically daily or operated in a batch fashion, with the data manually extracted weekly and sent by email.

Collected data included 12 coded attributes such as laboratory identification (Lab ID); engineered department (enDepart); code for patient type (e.g., OUT for outpatient); code for sex (e.g., M for male); age (e.g., 21); date (e.g., 02/01/2021); time (e.g., 23:04:06); instrument brand (e.g., Sysmex), and instrument identification (Instr ID), test name (e.g., AST for aspartate aminotransferase); testing result (e.g., 30), test unit (U/L). The laboratory can retrieve these attributes directly from the laboratory information system (LIS). The only requirement was for the laboratories to organize the data in a table according to the format below: "Lab ID; Depart; OUT; Sysmex; Instr ID; M; 21; 02/01/2021; 23:04:06; Sysmex; Instr ID; AST; 30; U/L" Being anonymized, the database was fully accessible to Beijing Center for Clinical Laboratories. Data were randomly selected from 11 hospitals at different levels, eight for modelling and three for independent external model testing. Data excluded criteria based on clinical rules were (1) lack of basic patient demographic information such as gender and age; (2) the most extreme values, roughly >4 SD away from the mean value in approximate normal distributions; (3) improper data format; (4) inconsistent units from different hospitals were unified; (5) non-patient materials (e.g., QC, research samples, dialysis fluids, animal samples, proficiency samples). Results were collected for 10 common analytes in serum, plasma or whole blood: white blood cell (WBC), red blood cell (RBC), hematocrit

(HCT), hemoglobin (HGB), platelet (PLT), aspartate aminotransferase (AST), alanine aminotransferase (ALT), glucose (GLU), total protein (TP), and albumin (ALB). These analytes were selected because they are the most frequently requested analyses. They also represent different result distributions commonly encountered in laboratory medicine.

### Error stimulation

Processed data of each analyte from eight hospitals (sequentially numbered one to eight) were divided as a training set and internal validation data with a ratio of training and internal validation sets of nine parts to one part for the three PBRTQC methods. In addition, the data from three hospitals (sequentially numbered from nine to 11) were used for three independent external testing sets. One thousand one hundred fifty data were allocated to each virtual day. The first 150 results were unbiased, and then a bias was introduced, starting from the 151 data points each day. The biased and unbiased data sets were dealt with in the same way. The original data collected represented unbiased data, and then corresponding biased data was produced by introducing a bias of different sizes according to formula (1):

$$x' = x \times (1 + P) \quad (P = -50\%, -48\%, -46\%, \dots, 46\%, 48\%, 50\%) \quad (1)$$

where  $x'$  represented the data after the bias was introduced,  $x$  was the original data, and  $P$  represented the specific relative bias value presented in a range from -50 to 50 % in steps of 2 %. Therefore, 50 biased data sets of different sizes were generated for each analyte based on unbiased data, covering common biases from real-world laboratory data sets.

### Additional information for mNL-PBRTQC

To identify causes of auto-correlation, eight independent variables, including hospital level, department type, patient type, sex, age, report date, report time and instrument brand, were selected.

Then the variables were re-codified into numerical formats. Patient types included outpatient, inpatient, emergency, and health check-up populations. Data collection periods were from January to March (set 1), April to June (set 2), July to September (set 3), and October to December (set 4). Report times were from 6:00 to 12:00 o'clock (set 1), 12:00 to 18:00 (set 2), 18:00 to 24:00 (set 3), and 24:00 to 6:00 of the next day (set 4). The minimum unit of age was one year. The age of patients was rounded. For example, 1.2 years would be counted as one year. Then the age value for each data would be added 1 year.

The department type was engineered to simplify the text information into numerical variables. A three-level scoring system was created for each feature based on the average value of the test results for tests where the keyword appeared. For example, if the average test value for keyword X is in the top 25 % of all keywords' average test value, then the sample with X in the department will be labelled as "1" for enDep. Similarly, samples with keywords in the middle 50 % will be labelled as "0". The keywords in the bottom 25 % will be labelled "-1" [22]. It was implemented in Python 3.9.12.

### mNL-PBRTQC algorithm

The ML-based Classification and Regression Tree (CART) algorithm was used for mNL-PBRTQC (Supplementary Figure 1). It is essential to create

a mapping relation  $f$ , whose function is to map each  $X_i$  to the corresponding  $Y_i$  accurately.  $X_i$  refers to the regression variables, and  $Y_i$  is the test result after the regression adjustment. The CART model was developed using the Python scikit-learn package, with all parameters set to default. The CART model was used to infer residuals, which was regarded as the input of our EWMA evaluation model, with predefined control limits (CLs) derived from optimal truncation limits (TLs) and block size (BS). The CL here referred to a reference range of each test item formulated from the processed data used for the judgement of quality control status. If the feature of data within its CL, represented in-control, if not represented out-of-control.

## Comparative study

The IFCC PBRTQC [10] and L-RARTQC [22] were used as comparative methods, while the mNL-PBRTQC was used as tests. The EWMA algorithm was used for model performance evaluation. As Bietenbeck et al. reported that winsorization and Box-Cox transformation improve the performance protocols, both of these steps were included at the beginning of the three QC methods.

The independent variables were sex, age, patient type, engineering department type, diagnosis, report date, report time, instrument brand and hospital level. The model labelled 5v included sex, age, patient type, engineered department type and engineered diagnosis; 4v included sex, age, patient type, and engineered department type; and 8v included sex, age, patient type, engineered department type, report date, report time, instrument brand and hospital level.

Firstly, the different models IFCC PBRTQC, 5vL-PBRTQC (Refers to the RARTQC based on statistical regression involving the five variations mentioned above), and 5vmNL-PBRTQC (Refers to the PBRTQC based on machine learning regression involving the five variations mentioned above) were compared using one hospital's data sets of six analytes for models differences; Secondly, these models IFCC PBRTQC, 4vmNL-PBRTQC and 8vmNL-PBRTQC (Refers to the PBRTQC based on machine learning regression involving the four and eight variations mentioned above) were compared by using eight hospitals' data for the same six analytes for observing the variable difference. Finally, IFCC PBRTQC, 4vL-RARTQC and 8vmNL-PBRTQC were used to comprehensively evaluate our mNL-PBRTQC. The general flow diagram of the experiment is shown in Figure 1.

## Evaluation metrics and implementation

The number of patient samples from the inception of the bias until error detection (NPed) was used to evaluate the performance of the three methods. The average and the median of NPeds (ANPed and MNPed) of minimal 600 virtual days in the training dataset and 20 virtual days in the test dataset served as performance metrics. The number of affected patient samples needed to detect the biases in 95 % of the simulations (95NPed) was also calculated. For each method and bias, an instability metric was determined based on all NPed of all days of the training set:

$$II = \text{Interquartile range NPed} / \text{median (NPed)}$$

When  $Pfr < 5\%$ , based on the overall MNPed, the lowest sum of MNPeds ( $\sum \text{MNPed}$ ) overall, biases for the three methods were selected. Similarly, the method with the lowest sum of 95NPeds ( $\sum 95\text{NPed}$ ) was selected. The lowest sum of I ( $\sum I$ ) was also considered. The "symmetric" CLs were set as a defaulted parameter. Infinite MNPeds or 95NPeds (when the error was not detected) were imputed with 1,100 (110 % of the maximum value).

$$\sum \text{MNPed} = \sum_{\text{bias}=-50\%}^{+50\%} \text{MNPed}_{\text{bias}}$$

$$\sum 95\text{NPed} = \sum_{\text{bias}=-50\%}^{+50\%} 95\text{NPed}_{\text{bias}}$$

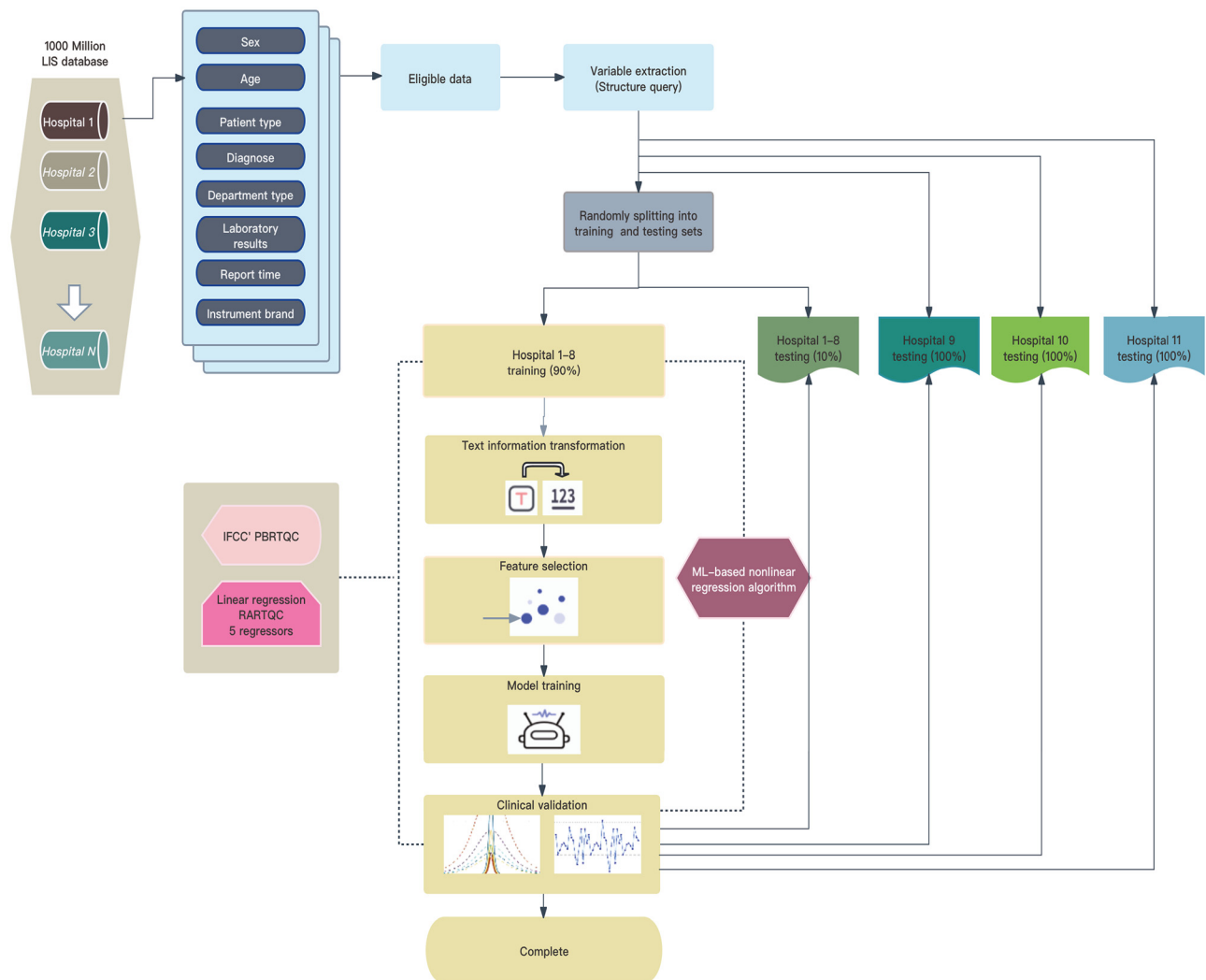
The implementation was performed in Python 3.9.12. The data processing, model training and testing process were based on "numpy", "pandas", "math", "random", "seaborn", "stats", and other tool kits. Error detection curve and validation charts were accomplished by matplotlib pyplot. The data was dealt with regression through Linear Regression and Decision Tree Regressor.

## Results

### Characteristics of baseline data

The data from eight hospitals, represented by five tertiary and three secondary hospitals, were used for model training. Data from three hospitals, representing two tertiary and one secondary hospital, was used for three independent model testing sets. The data from 10 analytes in 2021 were extracted from each hospital, including five routine blood analytes (WBC, RBC, HGB, PLT, HCT) and five biochemical measurands (ALB, TP, ALT, AST, GLU).

In the first step, a total of 8,046,680 whole blood routine data and 4,493,395 routine biochemical data were filtered by clinical rules. Then 6,786,110 whole blood routine data and 3,097,661 biochemical data were left in the training data set. The three test data sets were dealt with similarly (Supplementary Table 1). The distribution characteristics of each analyte for clinical-rule filtered data were descriptive as follows: the skewness of TP and HGB was 0.43 and  $-0.49$ , close to 0, representing near normal distribution; HCT, ALB, RBC and PLT were moderate skewness from  $-0.55$  to 2.12; GLU, WBC, ALT, AST demonstrated significant skewness from 15.2 to 57.46; HGB, HCT, ALB and RBC showed evident negative skewness; the kurtosis of the 10 test items ranged



**Figure 1:** General flow diagram of the experiment.

from 0.22 to 5,272; the three items with the largest skewness and kurtosis were AST, ALT and WBC (Supplementary Table 2).

In the second step, rule-filtered data were further processed using statistical or machine learning (ML) based manipulations, such as truncation limits (TLs) and BS optimization, Box-Cox transformation or not, and linear or nonlinear regression. The further processed data for the three methods had different performances in skewness and kurtosis parameters (Figure 2). The y-axis represented the distribution of the normalized data for the three methods.

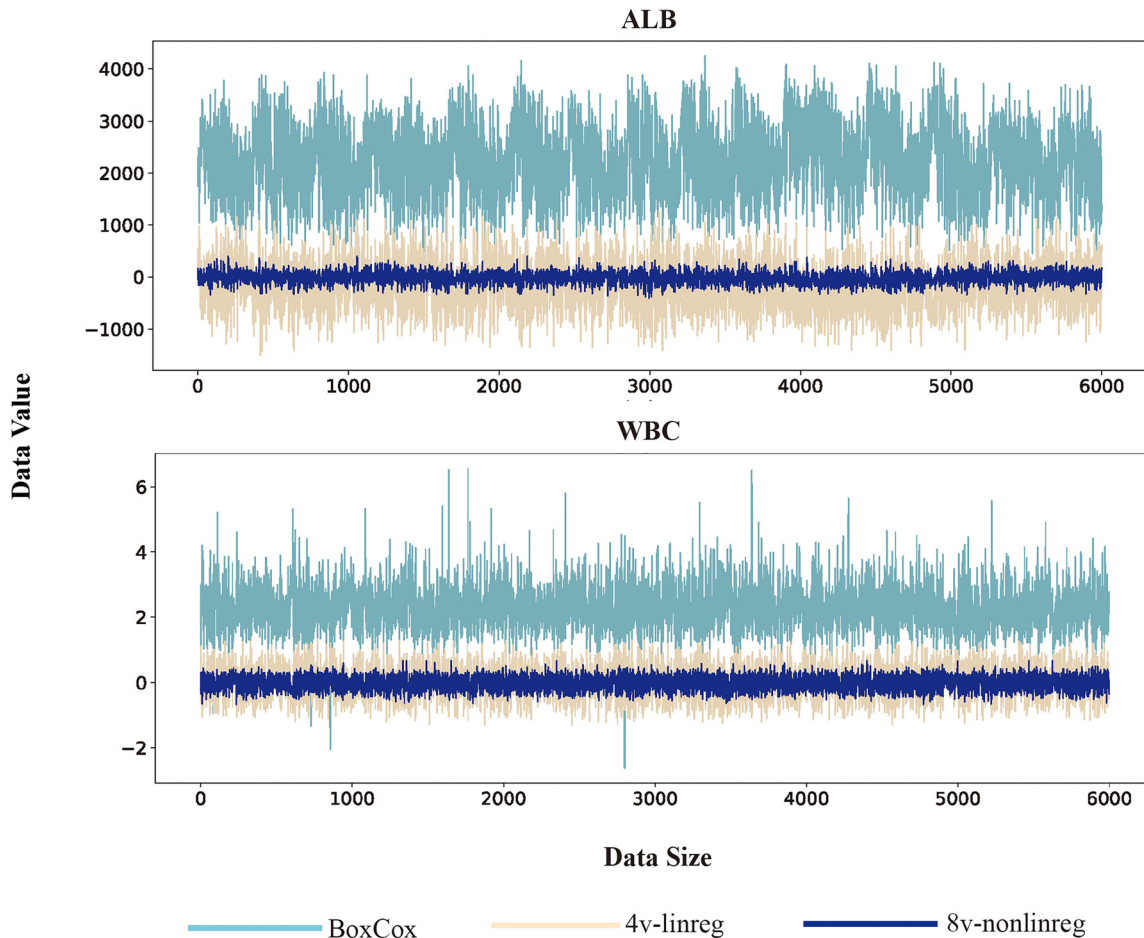
For ALB and WBC, the first 6,000 processed data for every method were selected for further investigation. The distribution of processed data for mNL-PBRTQC was more concentrated and stable than the other two PBRTQC methods

(Supplementary Table 3). The residual data obtained via regression present lower skewness and kurtosis. For ALB, the skewness values for IFCC PBRTQC, 4vL-RARTQC and 8vmNL-PBRTQC in sequence were  $-0.016271203$ ,  $0.192474425$  and  $-0.06228258$ . The kurtosis values for the three methods were  $-0.564349284$ ,  $-0.197640935$  and  $0.335691378$ . It indicated that 8vmNL-PBRTQC was better than the other two methods for the improvement of data concentration.

## Results of method comparison

Using the single tertiary hospital data, it was found that 5vmNL-PBRTQC and 5vL-RARTQC were superior and more symmetrical than the IFCC PBRTQC for all biases and all analytes (Supplementary Figure 2).





**Figure 2:** Trend of data for the three methods by using first 6000 processed data in 8 hospital training data set. The horizontal axis represented the number of data for analytes, the vertical axis represented the results processed for the three methods (by Box-Cox for IFCC PBRTQC; by 4v-linreg for L-RARTQC; by 8v-nonlinreg for mNL-PBRTQC).

## Results of variable comparison

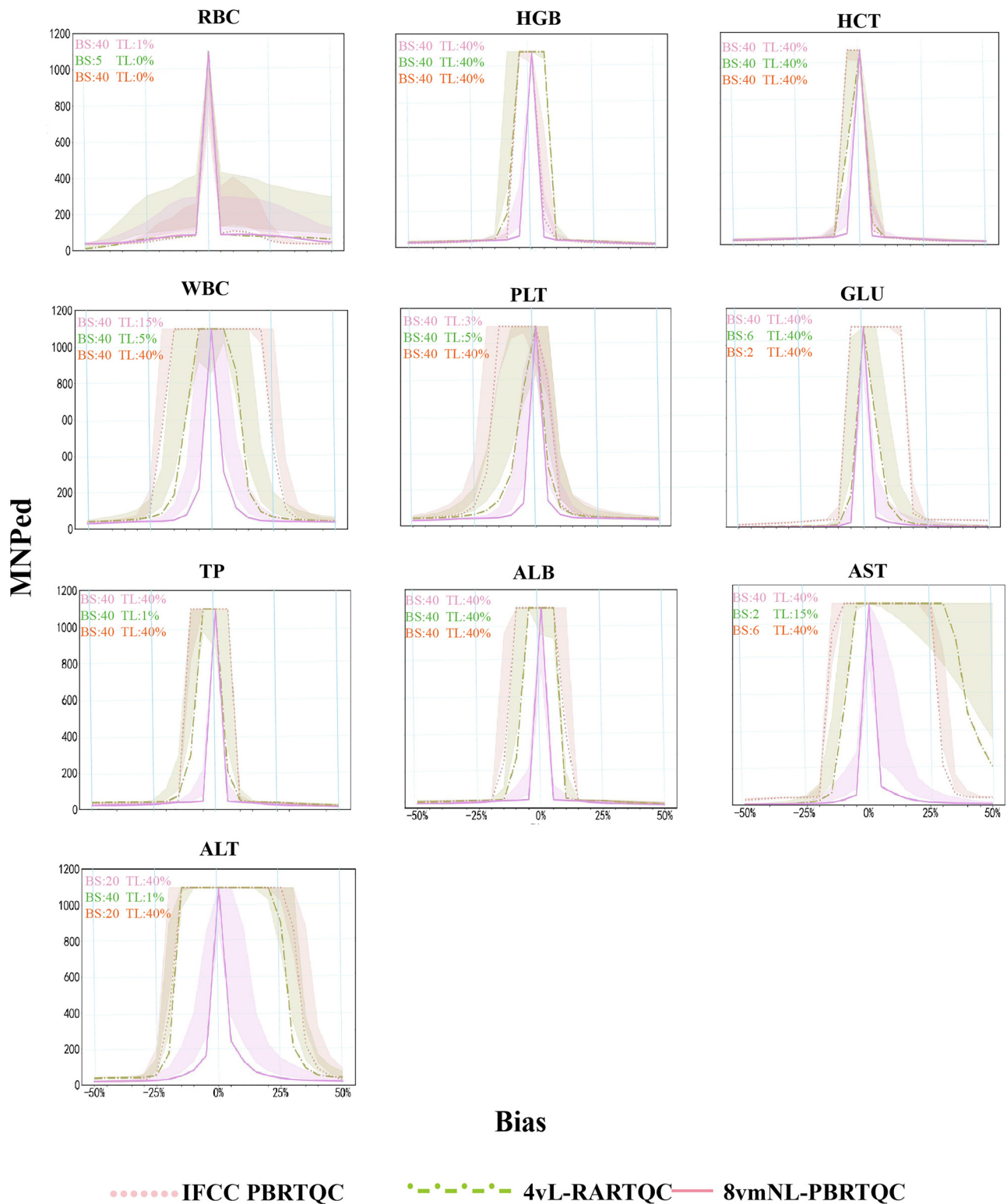
When the data set was expanded to all eight hospital data sets, it was found that mNL-PBRTQC with different regression variables was overall better than IFCC PBRTQC for all biases and all analytes. For WBC and ALT, 8vmNL-PBRTQC surpassed 4vmNL-PBRTQC in positive and negative directions for all biases (Supplementary Figure 3).

## Performance of IFCC PBRTQC, 4vL-RARTQC and 8vmNL-PBRTQC by using 11 hospital data

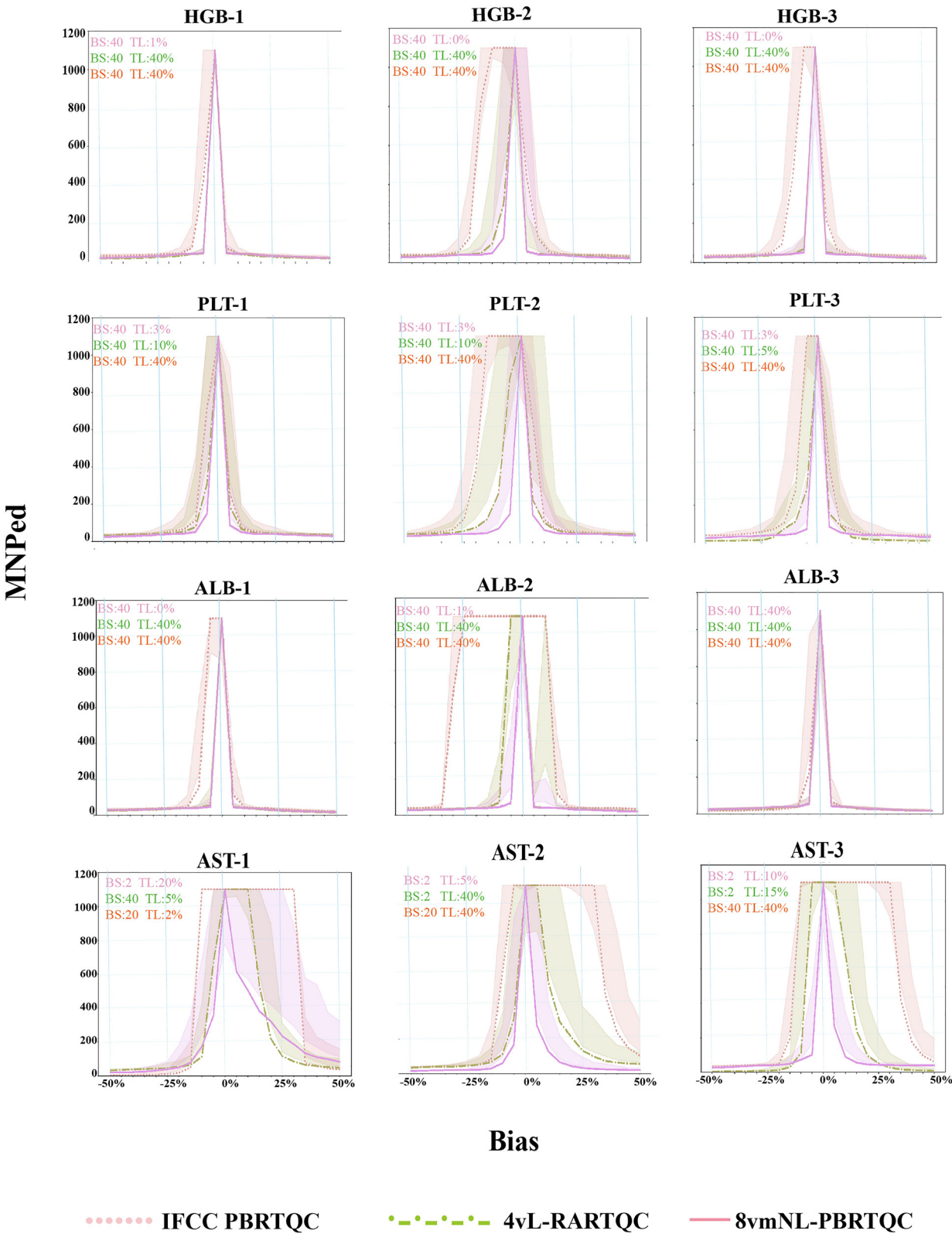
Using the training data set derived from the eight hospitals for all 10 analytes, the performance of IFCC PBRTQC, 4vL-RARTQC and 8vmNL-PBRTQC is shown (Figure 3 and Supplementary Table 4). The MNPeds of 8vmNL-PBRTQC were overall lower than the other two methods except for RBC

(2298 for IFCC PBRTQC, 2,415 for 4vL-RARTQC, and 2,460 8vmNL-PBRTQC). For an example of HGB,  $\pm 5\%$ ,  $-10\%$  and  $-15\%$  biases could be stably detected by 8vmNL-PBRTQC, while  $-5\%$  was not detected for IFCC PBRTQC and 4vL-RARTQC, and  $+5\%$  was not for 4vL-RARTQC.

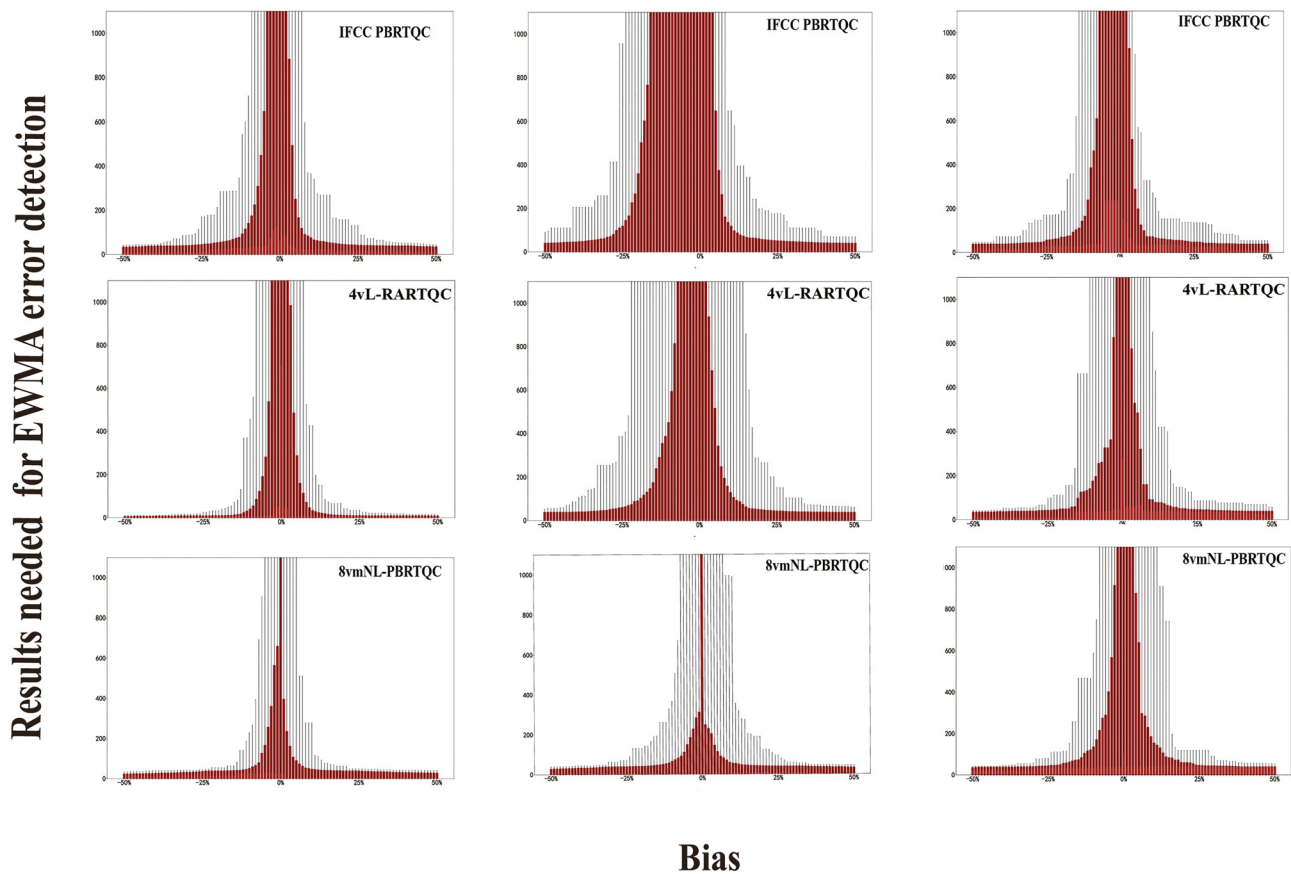
The results for the three independent test data sets from another three hospitals (sequentially numbered from 9 to 11) showed a similar change trend to the training data set (Figure 4 and Supplementary Tables 5–8). The overall performance of the three methods from best to least performance is 8vmNL-PBRTQC > 4vL-RARTQC > IFCC PBRTQC. In Hospital 9, the MNPed of 8vmNL-PBRTQC was lowest for PLT and AST and was equal to 4vL-RARTQC for HGB and ALB with 4vL-RARTQC, but no significant difference. In Hospital 10, the MNPed for 8vmNL-PBRTQC was the lowest for all analytes for all biases. In Hospital 11, the MNPed of 8vmNL-PBRTQC was the lowest for HGB, PLT, and AST, was equal to 4vL-RARTQC for ALB, and there was no significant difference.



**Figure 3:** Performance of IFCC PBRTQC, 4vL-RARTQC and 8vmNL-PBRTQC in the training data set from 8 hospital data. The horizontal axis represented the sizes of biases, the vertical axis represented MNPed for the three methods. In each diagram, the colored lines represented MNPed each bias, colored areas represented associated 95NPed. Optimal parameters were displayed in the top left corner (BS=block size; TL=truncation limit).



**Figure 4:** Performance of IFCC PBRTQC, 4vL-PRRTQC and 8vmNL-PBRTQC for 4 analytes in 3 independent test data sets from Hospital 9-11. The horizontal axis represented the sizes of biases, the vertical axis represented MNPeD for the three methods. In each diagram, the colored lines represented MNPeD each bias, colored areas represented associated 95NPed. Optimal parameters were displayed in the top left corner (BS=block size; TL=truncation limit). The number 1-3 in sequence represented Hospital 9-11.



**Figure 5:** Validation chart for IFCC PBRTQC and 4vL-RARTQC and 8vmNL-PBRTQC for PLT three independent test data sets from Hospital 9–11. Red bars represented the median number of results needed for error detection. The gray error bars represented the range of the number of results needed for error detection. Column from left to right represented Hospital 9–11.

Using PLT as an example, it was found that for a 20 % bias in either direction, the median (maximum) number of results needed for error detection (Y-axis) in the three test data sets (Figure 5), 8vmNL-PBRTQC was overall smaller than IFCC PBRTQC and 4vL-RARTQC. A +20 % bias was detected with 50 % probability (median) for the three hospitals, with 47–59 results needed for IFCC PBRTQC, 12–52 results for 4vL-RARTQC, and 40–61 results for 8vmNL-PBRTQC. A +20 % positive bias was detected in the three hospitals, which was 153–198 results for IFCC PBRTQC, 41–267 results for 4vL-RARTQC, and 62–121 results for 8vmNL-PBRTQC. A –20 % bias was detected with 50 % probability (median) for three hospitals was 55–361 results for IFCC PBRTQC, 11–93 results for 4vL-RARTQC, 40–58 results for 8vmNL-PBRTQC; –20 % bias was detected at the maximum of 189 patients for the IFCC PBRTQC, 32 results to not seen for 4vL-RARTQC, and 55–130 results for 8vmNL-PBRTQC.

## Discussion

PBRTQC refers to a statistical parameter (e.g., mean) that, over a defined number of patient results, is calculated to

monitor the performance of an analytical system. James O. Westgard states that one of the options for improving QC is to implement procedures that use patient data rather than depending on a few control measurements using traditional SQC (statistical quality control) procedures [26]. However, compared with conventional QC, PBRTQC is prone to be influenced by the measurement procedure, a range of pre-analytical, patient-related factors and other easily neglected variables. Thus, laboratory or instrument-specific PBRTQC pattern is found in the hospital-based setting due to the large variation of testing results [27]. The adaptability of PBRTQC is the core issue for the future implementation of these QC methods.

In our study, a crucial step is adopting ML-based nonlinear regression, called CART algorithm. Using one hospital data set, we found that two methods with regression step (5vmNL-PBRTQC and 5vL-RARTQC) were better for comparing methods with different working principles than the method without regression step (IFCC PBRTQC). When using the same five variables  $v$  for the two methods in the regression model, for most of analytes, 5vmNL-PBRTQC was



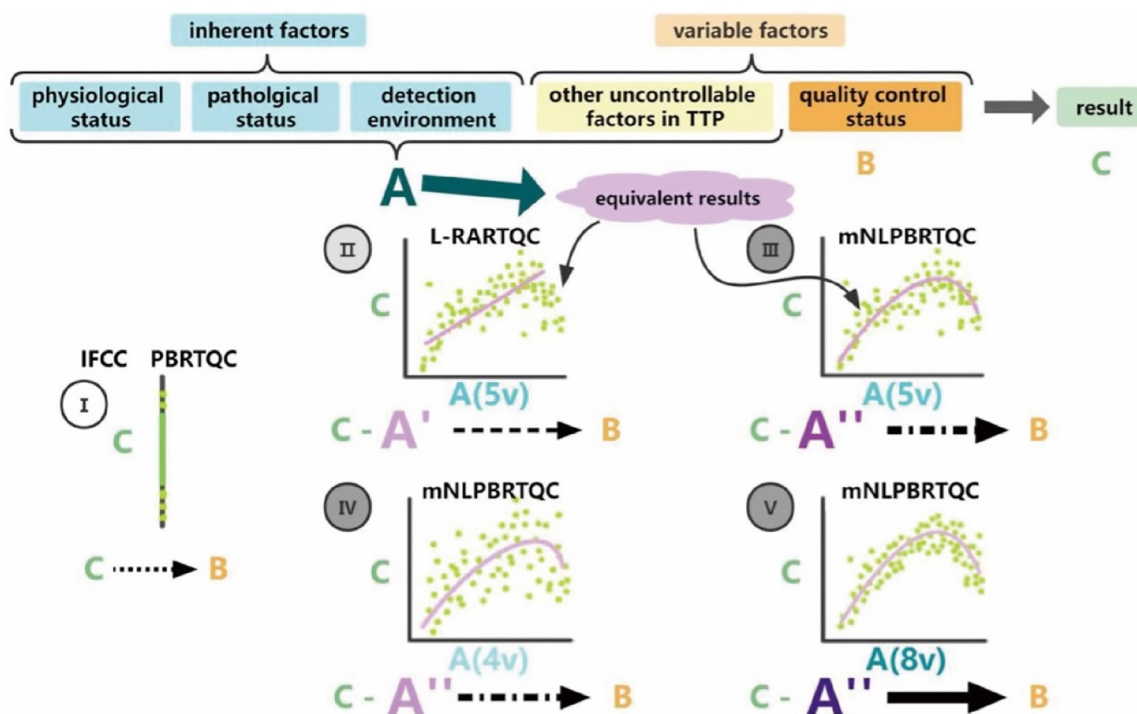
superior to 5vL-RARTQC for the accuracy and speed of error detection.

For biochemical measurands, ALT performed as the best one, with 5vmNL-PBRTQC's  $\sum \text{MNPed}$  down 8,527.5 and 5vL-RARTQC's  $\sum \text{MNPed}$  down 7,666; for routine blood analytes, WBC was best, with 5vmNL-PBRTQC's  $\sum \text{MNPed}$  down 3,127.5 and 5vL-RARTQC's  $\sum \text{MNPed}$  down 2,297. Therefore, compared with IFCC PBRTQC and 5vL-RARTQC, the performance of 5vmNL-PBRTQC is better, especially for those unsteady data and limited biases. This is because of the complex relationship between multiple variables and patient results. A simple linear model based on statistics limits the expression ability of information. Instead, a nonlinear model is closer to the essential distribution of variables and accurately expresses the relationship between multiple variables and patient results.

Comparing the 4vmNL- and 8vmNL-PBRTQC using a data set from eight hospitals, it was found that 8v- was better than 4v- for mNL-PBRTQC in positive and negative directions for all analytes and all biases. For ALT, the  $\sum \text{MNPed}$  at 8v was reduced by 954 compared to at 4v; for GLU, it was reduced by 382. The reason is that along with the expansion of a data set equal to the increase of relative instability of data for each analyte, 8v represents expanded data better than 4v. In addition, it was found that it is important not to

include diagnostic information in the model as a variable because (1) the difference in physician skill and laboratory testing capacity, e.g., for the same disease, in a primary clinic, doctors may be given descriptive word in the part of diagnostic information column, but in a tertiary hospital, doctors will give a definite diagnosis. (2) Missing or indefinite diagnostic information for outpatients (3) mNL-PBRTQC used for error detection ahead of doctor's diagnosis.

The model's performance surpassed the reported methods, and three independent test data sets showed similar applicability. This can be explained from the perspective of information entropy. Many factors affect patient results (Figure 6). The patient results are labelled as C, QC status as B, and other factors exclusive of QC status as A. The IFCC PBRTQC is inferred B by C, while C is expressed by (I), only with one dimension. The L-RARTQC is equivalent to introducing A to C, expanding C to multiple dimensions. The core of the regression step is equal to converting A's information to C's information with the same unit. Consider that the smaller the fitting residual value of A, the better the error detection performance. Here (I), (II) and (III) represent different methods, mNL-PBRTQC adopted a nonlinear regression with a smaller fitting residual value than L-RARTQC, so the performance for mNL-PBRTQC is better than L-RARTQC. While (I), (IV), and



**Figure 6:** Algorithmic interpretation based on information entropy. (I), (II) and (III) represented the comparison of different methods; (I), (IV), (V) represented the comparison of a different number of variables (v). The color depth of letters, the thickness of lines and the degree of line of continuity all indicated the amount of information.



(V) represent different variables. As the number of variables increases, the equivalent value of A is closer to the nature of C. C becomes less discrete after the mapping, with a smaller fitting residual value. The overall performance of the three methods from best to least: 8vmNL-PBRTQC > 4vL-RARTQC > IFCC PBRTQC.

The proposed approach can merge multiple laboratories to jointly establish the method during the application stage. It [23] has reported the merging of four measurement sets to enhance the universality of the approach. Merging data from different sources would increase the width of the

distribution. Since the introduction of residuals can narrow the width of the controlled variable distribution through baseline adjustment, the distribution width of residuals would increase when the width of the original data distribution widens, thereby leading to widening of the control limit. During the method construction, we merged eight hospitals' data to achieve universality. By comparing the mean and standard deviation of the eight hospitals' data, it was ensured that the increase of width in residual distribution would not exceed it in any of the data sources. This condition was the prerequisite for hospital merging.

The model has been deployed in the LIS of another two different hospitals. Some specific monitoring scan-shots from LIS of one hospital are shown (Figure 7). The interface demonstrated in Figure 7A shows that, compared to conventional QC, the performance of mNL-PBRTQC is equal to or better than conventional QC for the same patient testing results. The interface shown in Figure 7B shows that there is also suitable for different instruments for monitoring the same analytes, even for other testing sites.

**Acknowledgments:** We thank all those who participated in this study.

**Research ethics:** The project was approved by the local hospital Ethics Committee.

**Informed consent:** Not applicable.

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

**Competing interests:** Authors state no conflict of interest.

**Research funding:** This work was supported by Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (ZYLX201811), Excellence project of key clinical specialty in Beijing and National Natural Science Foundation of China (72374145).

**Data availability:** The data are not publicly available.

## References

1. 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.
2. 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.
3. 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.
4. Miller WG, Ereik 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.
5. Algeciras-Schimmich 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.
6. 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.
7. Thaler MA, Iakoubov 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.
8. 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* 2018;28:010705.
9. Loh TP, van Rossum HH, Katayev A, Cervinski MA, Bietenbeck A, Badrick T. Patient-based real-time quality control: review and recommendations. *Clin Chem* 2019;65:962–71.
10. Bietenbeck A, Cervinski MA, Katayev A, Loh TP, van Rossum HH, Badrick T. Understanding patient-based real-time quality control using simulation modeling. *Clin Chem* 2020;66:1072–83.
11. Hoffmann RG, Waid ME. The “average of normals” method of quality control. *Am J Clin Pathol* 1965;43:134–41.
12. Duan X, Wang B, Zhu J, Shao W, Wang H, Shen J, et al. Assessment of patient-based real-time quality control algorithm performance on different types of analytical error. *Clin Chim Acta* 2020;511:329–35.
13. 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.
14. Neubauer AS. The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation. *Clin Chem* 1997;43:594–601.
15. Linnet K. The exponentially weighted moving average (EWMA) rule compared with traditionally used quality control rules. *Clin Chem Lab Med* 2006;44:396–9.
16. 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.
17. Jones GR. Average of delta: a new quality control tool for clinical laboratories. *Ann Clin Biochem* 2016;53:133–40.
18. Zhou R, Wang W, Padoan A, Wang Z, Feng X, Han Z, et al. Traceable machine learning real-time quality control based on patient data. *Clin Chem Lab Med* 2022;60:1998–2004.
19. van Rossum HH. Moving average quality control: principles, practical application and future perspectives. *Clin Chem Lab Med* 2019;57:773–82.
20. Zhou R, Liang YF, Cheng HL, Padoan A, Wang Z, Feng X, et al. A multi-model fusion algorithm as a real-time quality control tool for small shift detection. *Comput Biol Med* 2022;148:105866.
21. Liang Y, Wang Z, Huang D, Wang W, Feng X, Han Z, et al. A study on quality control using delta data with machine learning technique. *Heliyon* 2022;8:e09935.
22. Zhou Q, Loh TP, Badrick T, Lim CY. Impact of combining data from multiple instruments on performance of patient-based real-time quality control. *Biochem Med* 2021;31:020705.
23. Duan X, Wang B, Zhu J, Zhang C, Jiang W, Zhou J, et al. Regression-adjusted real-time quality control. *Clin Chem* 2021;67:1342–50.
24. Loh TP, Bietenbeck A, Cervinski MA, van Rossum HH, Katayev A, Badrick T. Recommendation for performance verification of patient-based real-time quality control. *Clin Chem Lab Med* 2020;58:1205–13.
25. Loh TP, Cervinski MA, Katayev A, Bietenbeck A, van Rossum H, Badrick T. Recommendations for laboratory informatics specifications needed for the application of patient-based real time quality control. *Clin Chim Acta* 2019;495:625–9.
26. Westgard JO, Bayat H, Westgard S. Advanced QC strategies: risk-based design for medical laboratories, 1st ed 7614 Gray Fox Trail Madison WI 53717: Westgard Quality Corporation; 2022:131 p.
27. 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.

**Supplementary Material:** This article contains supplementary material (<https://doi.org/10.1515/cclm-2023-0964>).