Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A multi-model fusion algorithm as a real-time quality control tool for small shift detection

Rui Zhou [a,b,1], Yu-fang Liang [a,1], Hua-li Cheng [a], Andrea Padoan [c], Zhe Wang [d], Xiang Feng [d], Ze-wen Han [d], Biao Song [d], Wei Wang [e], Mario Plebani [c,*], Qing-tao Wang [a,b,**]

[a] Department of Laboratory Medicine, Beijing Chao-yang Hospital, Capital Medical University, Beijing, PR China
[b] Beijing Center for Clinical Laboratories, Beijing, PR China
[c] Department of Laboratory Medicine, University-Hospital of Padova, 35128 Padova, Italy
[d] Inner Mongolia Wesure Date Technology Co., Ltd, Inner Mongolia, PR China
[e] Department of Blood Transfusion, Beijing Ditan Hospital, Capital Medical University, Beijing, PR China

ARTICLE INFO

Keywords:
Quality control
PBRTQC
Small shift
Machine learning
Information entropy

ABSTRACT

Background: Patient-based real-time quality control (PBRTQC), a complement to traditional QC, may eliminate matrix effect from QC materials, realize real-time monitoring as well as cut costs. However, the accuracy of PBRTQC has not been satisfactory as physicians expect till now. Our aim is to set up a artificial intelligence-based QC for small error detection in real laboratory settings. Taking tPSA as our unique research subject, data extraction, data stimulation, data partition, model construction and evaluation were designed.

Methods: 84241 deidentified results for tPSA were extracted from Laboratory Information System of Aviation General Hospital. The data set was accumulated by way of data simulation. Independent training and test datasets were separated. After three classification models (RF, SVM and DNN) in ML constructed and weighted by information entropy, a multi-model fusion algorithm was generated. Performance of the fusion model was evaluated by comparing with optimal PBRTQC.

Results: For 4 PBRTQC methods, MovSO showed overall better performance for 0.2 μg/L bias and optimal MNPed was equal to 200. For the fusion model, MNPeds were less than 12 for all biases, and ACC surpassed MovSO nearly 100 times. Except for 0.01 μg/L bias, ACC was more than 0.9 for the rest of biases. FPR was apparently lower than MovSO, only 0.2% and 0.1%.

Conclusion: The fusion model shows outstanding performance and reduces incorrect and omitting error detection, adaptable for the real settings.

1. Introduction

Prostate cancer in male is the most frequently diagnosed cancer and is the second one with regard to mortality according to World Health Organization (WHO) Report in 2019 [1], and ranks 6th in the crude incidence of male malignant tumors in China [2]. Prostate specific antigen (PSA)testing is crucial for the diagnosis, monitoring and treatment of prostate cancer [3]. A recent study reported that a <0.2 μg/L change in serial PSA measurements may regard as indicative of poorer prognosis [4]. Therefore, the detection for small shifts, like PSA analyte, is not ignored in patient management [5].

Patient-based real-time quality control (PBRTQC) as a dynamic QC tool, directly implements QC monitoring by real patient results, thus not only eliminating matrix effect of traditional QC materials but cutting the total cost. Related studies have emerged in large numbers in recent years. Moving average (MA), moving median (MM), moving quartile (MQ), and exponentially weighted moving average (EWMA), EWMA [6, 7] showed outstanding performance for the detection of systematical error (SE), simulated annealing (SA) algorithm–developed MA protocols enabled to rapidly detect SE compared with MA [8]. The moving

[*] Corresponding author. Department of Laboratory Medicine, University-Hospital of Padova, 35128, Padova, Italy.

[**] Corresponding author. Department of Laboratory Medicine, Beijing Chao-yang Hospital, Capital Medical University, Beijing Center for Clinical Laboratories, No.8 Gongti South Street, Chaoyang District, Beijing, 100020, PR China.

E-mail addresses: Mario.plebani@unipd.it (M. Plebani), wqt36@163.com (Q.-t. Wang).

[1] Rui Zhou and Yu-fang Liang contributed equally as co-first authors to this work.

https://doi.org/10.1016/j.compbiomed.2022.105866
Received 16 May 2022; Received in revised form 23 June 2022; Accepted 9 July 2022
Available online 12 July 2022

standard deviation (MovSD) and the moving sum of number of patient results (MovSO) exceeded other statistical algorithms to the detection of random error or small shift [4]. In addition, MM, MovSO, or MQ, showed more robust than MA for skewed distribution data [7]. However, the accuracy of PBRTQC has not been satisfactory till now, especially for tPSA analyte of this kind.

Machine learning (ML), a branch of artificial intelligence (AI), different from traditional statistics, with the characteristics of high accuracy, is commonly used for disease prediction, and automatic decision-making in medical field., For example, computer vision, speech recognition and natural language processing algorithms are applied to image, audio, and text files in the medical field, respectively.

Our aim is to set up an artificial intelligence-based QC for small error detection in real laboratory settings. Taking PSA as our unique research subject, data extraction, data simulation, data partition, model construction and evaluation were designed.

2. Materials and methods

2.1. ML QC

2.1.1. Data simulation

84 241 deidentified results for tPSA analyte measured on Abbott I2000 (SR, Chicago, USA) from Jan 2016 to Jun 2021, were extracted from the laboratory information system, Aviation General Hospital. Original matched tPSA reagent and calibrator from Abbott Diagnostics Division, and commercial QC materials from BIO-RAD manufacturer were tested. All procedures strictly followed ARCHITECT' instruction. The Westgard $1_{3S}/2_{2s}$ quality rules were selected and 1% outliers for all patient results were excluded. Preliminary, about half of the data were used for the training dataset and the latter for test dataset, as shown in Fig. 1.

The population $\bar{x}$ and CV_t of the first group of subjects were calculated by 43 699 training data. Data augmentation was performed according to IFCC method [6] in 43 699 training data, the random number in the range of (−0.3–0.3)μg/L added to each data, 43 699 new data was obtained. After 80 rounds of data analogy, finally a total 3 495 920 data was accumulated, which was assumed to be error-free, called positive data in ML. The $\bar{x}$ and CV_t of the current data set were validated to be like the original data set. Both sizes of 0.02 μg/L and 0.15 μg/L biases were introduced to the 3 495 920 positive data to build up the simulated negative data for training dataset. Therefore, two training models of 0.02 μg/L and 0.15 μg/L biases was constructed. While different sizes of 0.01 μg/L、0.03 μg/L 、0.05 μg/L、0.08 μg/L、0.10 μg/L、0.20 μg/L biases were introduced for test dataset for validation. Here 0.10 μg/L bias was roughly equal to the absolute value of the critical systematical

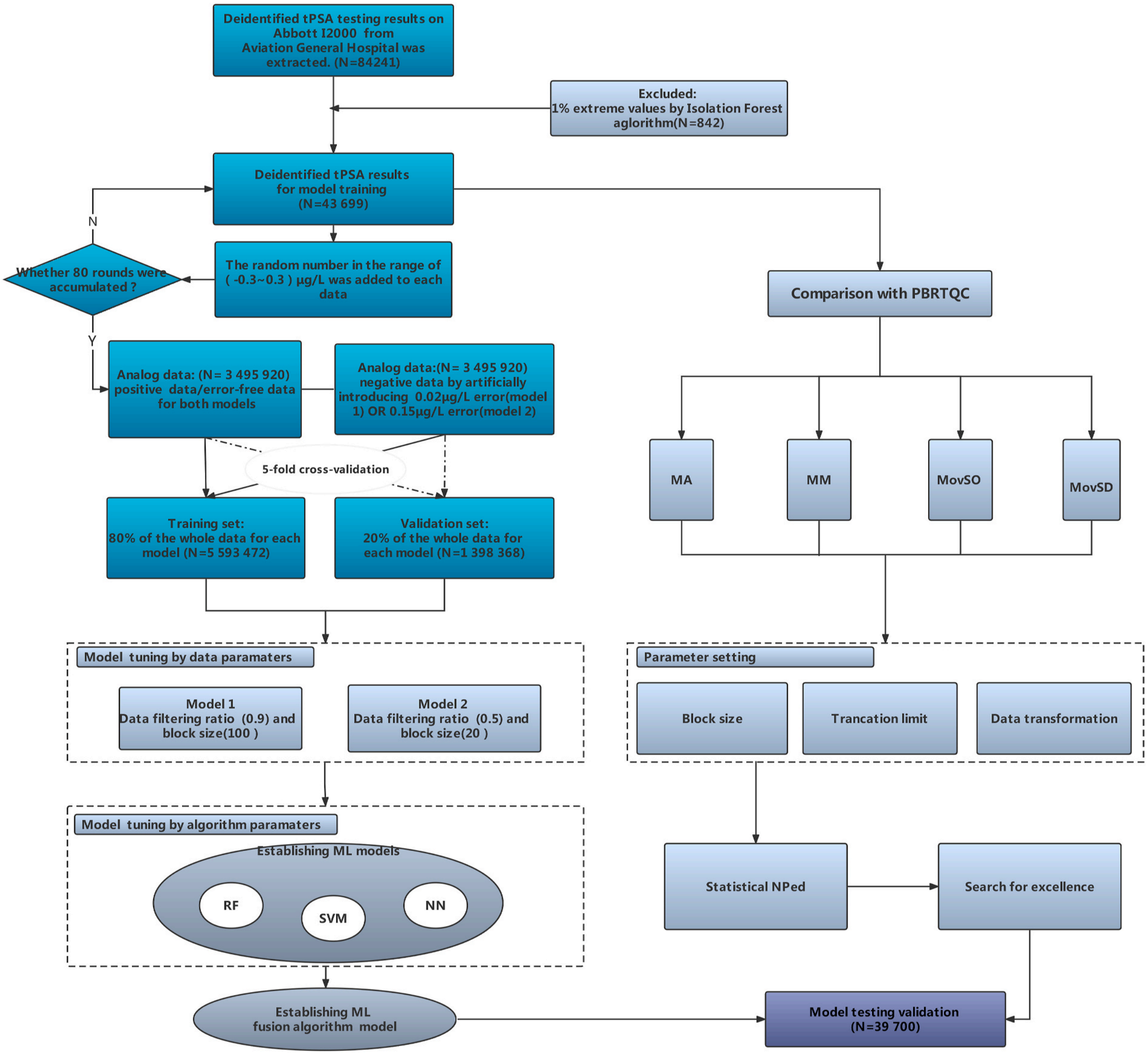
Deidentified tPSA testing results on Abbott I2000 from Aviation General Hospital was extracted. (N=84241)
Excluded: 1% extreme values by Isolation Forest aglorithm(N=842)
Deidentified tPSA results for model training (N=43 699)
N
Whether 80 rounds were accumulated ?
The random number in the range of (-0.3~0.3) μg/L was added to each data
Y
Analog data: (N= 3 495 920) positive data/error-free data for both models
Analog data:(N= 3 495 920) negative data by artificially introducing 0.02μg/L error(model 1) OR 0.15μg/L error(model 2)
5-fold cross-validation
Training set: 80% of the whole data for each model (N=5 593 472)
Validation set: 20% of the whole data for each model (N=1 398 368)
Model tuning by data paramaters
Model 1 Data filtering ratio (0.9) and block size(100)
Model 2 Data filtering ratio (0.5) and block size(20)
Model tuning by algorithm paramaters
Establishing ML models
RF
SVM
NN
Establishing ML fusion algorithm model
Comparison with PBRTQC
MA
MM
MovSO
MovSD
Parameter setting
Block size
Trancation limit
Data transformation
Statistical NPed
Search for excellence
Model testing validation (N=39 700)

Fig. 1. The flowchart diagram of MLQC for small shift detection.

bias at the 0.53 μg/L level for tPSA. In the paper, a bias represented a shift in the mean.

2.1.2. Information entropy model construction

Firstly, 84 241 raw data were pre-processed. This process mainly included data filtering, and data normalization. Data filtering was done by isolated forest algorithm, to remove outliers. Data normalization was performed for data scaling. Secondly, block size should be pre-defined starting successive experiment. The "block size" indicates the number of patient results that are included in each calculation of MLQC or PBRTQC. We also can consider it as a key tuning parameter of all algorithms. Random Forest (RF) was used to identify the optimal block size of each algorithm using pre-set default values(number of trees = 200, max tree depth = 200). Bias-free data and 0.10 μg/L biased data were used. Block sizes from 5 to 15 in steps of 1 were measured. Thirdly, Random Forests (RF), Support Vector Machine (SVM), and Neutral Network (NN) algorithms were recruited and trained based on predefined block size, Fig. 2D–F. The data of a block size was defined as a new "ML sample". Bias-free data was labeled 0, and biased data was 1. In order to ensure the stability and optimization of single algorithm, 5-fold cross validation was adopted. Specifically, the training data set was further randomly divided into five parts, four of which were used as training data in turn, and one of which was used as verification data independently, so as to repeat the experiment above five times. The corresponding accuracy was obtained each time. The average of accuracy five times was used to estimate the accuracy of the algorithm.

In order to further improve prediction ability of our model for small shift, information entropy fusion algorithm was introduced to the three optimal models above mentioned, Fig. 2G. Information entropy here was for determining weight factors of each single model. The experiment in detail was as followed:

Because of output value of our classification models (RF/SVM/NN) labeled 0 or 1, in order to obtain continuous error value, we directly extracted the corresponding probability value of prediction error *e* from each label. To calculate specific proportion of relative error for a single model, the sum of absolute errors for all samples was first calculated, It was calculated as follows:

$$p_{ji} = \frac{e_{ji}}{\sum_{i=1}^{n} e_{ji}}, i = 1, 2, ..., n$$

e_{ji} was the prediction error of the *j*-th prediction model for the *i*-th test sample, p_{ji} was the proportion of prediction relative error of the j-th prediction model to the i-th similar sample, $\sum_{i=1}^{n} p_{ji} = 1$, the *j* was the number of single prediction model, $j = 1,2, ...,k$.

The entropy of the relative error of the *j*-th prediction model was calculated, the formula was

$$H_j = -\sum_{i=1}^{n} p_{ji} \log_2 p_{ji}, j = 1, 2, ..., k$$

The variation coefficient D_j of the prediction relative error of the j-th prediction model was calculated, the formular was

$$D_j = 1 - H_j$$

The weight coefficient of the j-th prediction model was calculated, the formula was

$$W_j = \frac{1}{k-1}\left(1 - \frac{D_j}{\sum_{j=1}^{k} D_j}\right)$$

W_j was the weight factor of the j-th model, $\sum_{i=1}^{k} W_j = 1$. If the prediction error of a single model was smaller, the entropy value of the model was larger, indicating that the model was more stable, then the weight of the model was larger in our combined model. On the contrary, if the information entropy of a model was relatively smaller, it indicated that the model was unstable, then the weight in the combined model was smaller.

2.2. Patient-based real-time quality control (PBRTQC)

In our work, the following PBRTQC algorithms, MA, MM, MovSD and

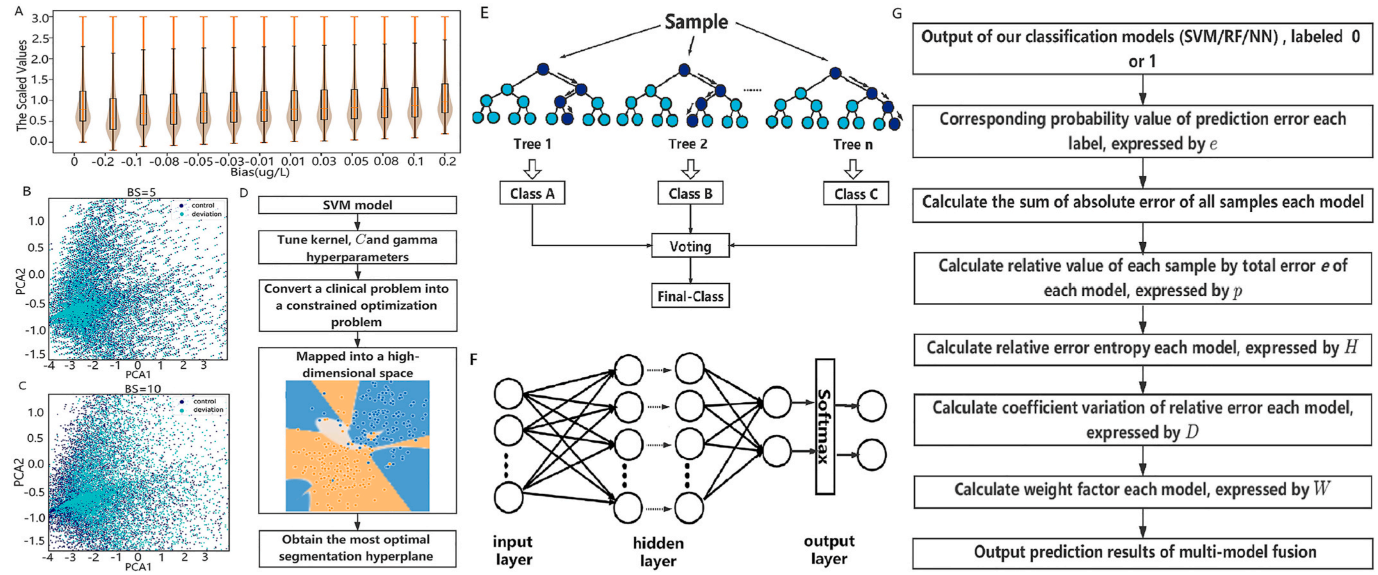
A
The Scaled Values
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
-0.2
-0.1
-0.08
-0.05
-0.03
-0.01
0.01
0.03
0.05
0.08
0.1
0.2
Bias(ug/L)
B
BS=5
control
deviation
PCA2
PCA1
C
BS=10
D
SVM model
Tune kernel, C and gamma hyperparameters
Convert a clinical problem into a constrained optimization problem
Mapped into a high-dimensional space
Obtain the most optimal segmentation hyperplane
E
Sample
Tree 1
Tree 2
Tree n
Class A
Class B
Class C
Voting
Final-Class
F
input layer
hidden layer
output layer
Softmax
G
Output of our classification models (SVM/RF/NN) , labeled 0 or 1
Corresponding probability value of prediction error each label, expressed by e
Calculate the sum of absolute error of all samples each model
Calculate relative value of each sample by total error e of each model, expressed by p
Calculate relative error entropy each model, expressed by H
Calculate coefficient variation of relative error each model, expressed by D
Calculate weight factor each model, expressed by W
Output prediction results of multi-model fusion

Fig. 2. Data visualization and working principles of ML algorithms.
A. Data distribution features of unbiased data and biased data with different sizes;
B–C. Visualization of two kinds of data by PCA technique for 5 and 10 block sizes when introducing 0.15 μg/L bias.
D. Illustration diagram of SVM model.
E. Illustration diagram of the RF model.
F. Illustration diagram of NN model.
G. Working principal diagram of multi-model fusion model.

MovSO were selected as reference QC approaches [6]. MA and MovSD were calculated on neat, and BOX-COX transformed test results.

Truncation limits (TLs), were experimented to exclude of 10%、7.5%、5%、3%、2.5%、2%、1.5%、1%、0.5% or 0% on each tail of data distribution for MA, MM and MovSD [4,7], Here, Winsorization method replaced outlying values with the corresponding lower and upper TL(LTL, UTL) that were exceeded [6], expressed by the formula as followed.

$$\omega(x)=\begin{cases} LTL\ if\ x\ < LTL \\ UTL\ if\ x > UTL \end{cases}$$

Block sizes (BSs) of 20、50、100、200、500、1000 and three methods for control limits (CLs) calculation were evaluated as IFCC's approach for 4 algorithms [4]. The combination of TLs, BSs, CLs and algorithms were tested as the above mentioned.

For MovSO method, tPSA concentration < 0.4 μg/L was considered as representing a "normal" population. The tPSA results were converted in to binary status ("0" = tPSA below detection limit, or "1" = tPSA above detection limit). The moving sum of number of positive results was used in statistics in our study [4].

All results with 1% extreme values excluded initial our experiment were kept in their original order and divided into 20 visual days for training dataset and for test dataset separately with 2000 measurements each day. For both datasets, assuming data error-free, we started from the measurement in the original order, smoothed toward the end in a step of 1, collected 1000 queues each block size (20、50、100、200、500、1000) for each virtual day, used for the calculation of false positive rate(FPR) and true negative rate(TNR). When an artificial error was added, starting from the last result of the first block of results in the original order, we smoothed in a step of 1, also collected 1000 queues each block size for each virtual day, used for the calculation of false negative rate(FNR), true positive rate(TPR), ANPed, MNPed, 95NPed. Sustained positive biases of 0.01 μg/L、0.02 μg/L、0.03 μg/L、0.05 μg/L、0.08 μg/L、0.10 μg/L、0.15 μg/L、0.20 μg/L was introduced. We recorded the first tPSA result that exceeded the CL.

The average values of FPR, TNR, FNR, TPR, ANPed, MNPed, 95NPed each bias for 20 visual days on both datasets were calculated. Additionally, ANPed at critical systematic bias was calculated. When FPR was no more than 0.05, the method with the lowest sum of MNPeds and the highest average TPR over all biases was selected. The method with the lowest sum of ANPed also considered.

2.3. Evaluation criteria

We use area under the receiver operating curve (AUC), Accuracy (ACC), true positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) were used for evaluating the analytical performance of algorithms. The process parameters of true positive(TP), true negative(TN), false positive(FP) and false negative (FN) were recorded. They were expressed as followed in formula and Table 1:

When FPR ≤ 0.05, the number of patient samples from the inception of the bias until error detection (NPed) was evaluated for all methods. The average, the median and 95 quartile of NPeds (ANPed, MNPed and 95NPed) of all 20 virtual days of test dataset served as performance metrics. ML model analysis was implemented in Python 3.7.3. All software packages were accessed from the sklearn library_2.4.0 in the public Python.

3. Results

3.1. Data description

43699 pre-processed data were obtained, which were expanded to 3495920 data through data simulation as training dataset. For unbiased data, the kurtosis was 219.99 and the skewness was 12.47. After twelve biases introduced to generate the corresponding biased data, unbiased and biased data distribution were shown in Fig. 2A. Among them, 0 bias corresponded to unbiased data. To facilitate observation, all data were scaled to 0–3. The data was concentrated near the small value as a whole, and the outliers were distributed at the end of the maximum value; With the increase of the bias introduced from left to right on x-axis, the data distribution was present at a linear shift as a whole, but the kurtosis and skewness of biased data in each group had no significant change compared with unbiased data; the overlapping between biased data and unbiased data occurred dramatically for all biases, therefore, it has been proved that the data was poorly separable in two dimensions.

Table 1
The definition of confusion matrix parameters.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Confusion matrix		Identified type	
		Biased data	Biased data
True type	Biased data	TP	FN
	Unbiased data	FP	TN

3.2. Performance for information entropy algorithm

The block size of each single model was set to 10. Because of the AUC of different block sizes in RF observed, when a block size was less than 10, the AUC increased by 0.23 for each increase of block size on average; but when a block size was greater than 10, the change of AUC value was no longer significant. Additionally, it was obvious that the larger the block size was, the stronger the data prone to be separable, seen in Fig. 2B and C.

Because single algorithm performed differently in the same scenario, we designed a MLQC framework by weight of information entropy to each independent algorithm to make up the shortcomings of each algorithm and to adopt for real and complex laboratory data. An analysis of Information entropy could measure the randomness of a probability distribution, referring to the size of information contained. As to PSA for all biases, the performance of each algorithm was also different. In consideration with each model adjusted by the output loss function, the weighting the output probability of the model was equivalent to smoothing the loss function of ML, thus reducing the overall output error and improving the accuracy of our QC model. Through an analysis by way of information entropy in our study, the weight factor of our classifier models (SVM, RF and NN) was given to 0.34, 0.27 and 0.39 respectively, in order to adjust prediction outputs of our fusion model. As shown in Fig. 3A-E, for all biases, the ACC, TPRs and TNRs of the fusion model were better than those of the three single models, and the FPR and FNR were the lowest among all ML models. Compared with SVM, RF and NN, the accuracy of the fusion model was improved by 8.7%, 9.6% and 6.9% respectively, in Fig. 3E. Affected patient samples before error detection of the four ML models were shown in Fig. 3F, the MNPed on average for RF、SVM、NN and fusion Algorithm was 7.17, 7.25, 9.25 and 5.5. It proved that fusion algorithm for detecting bias of different sizes was faster than that of three single algorithm, leading less

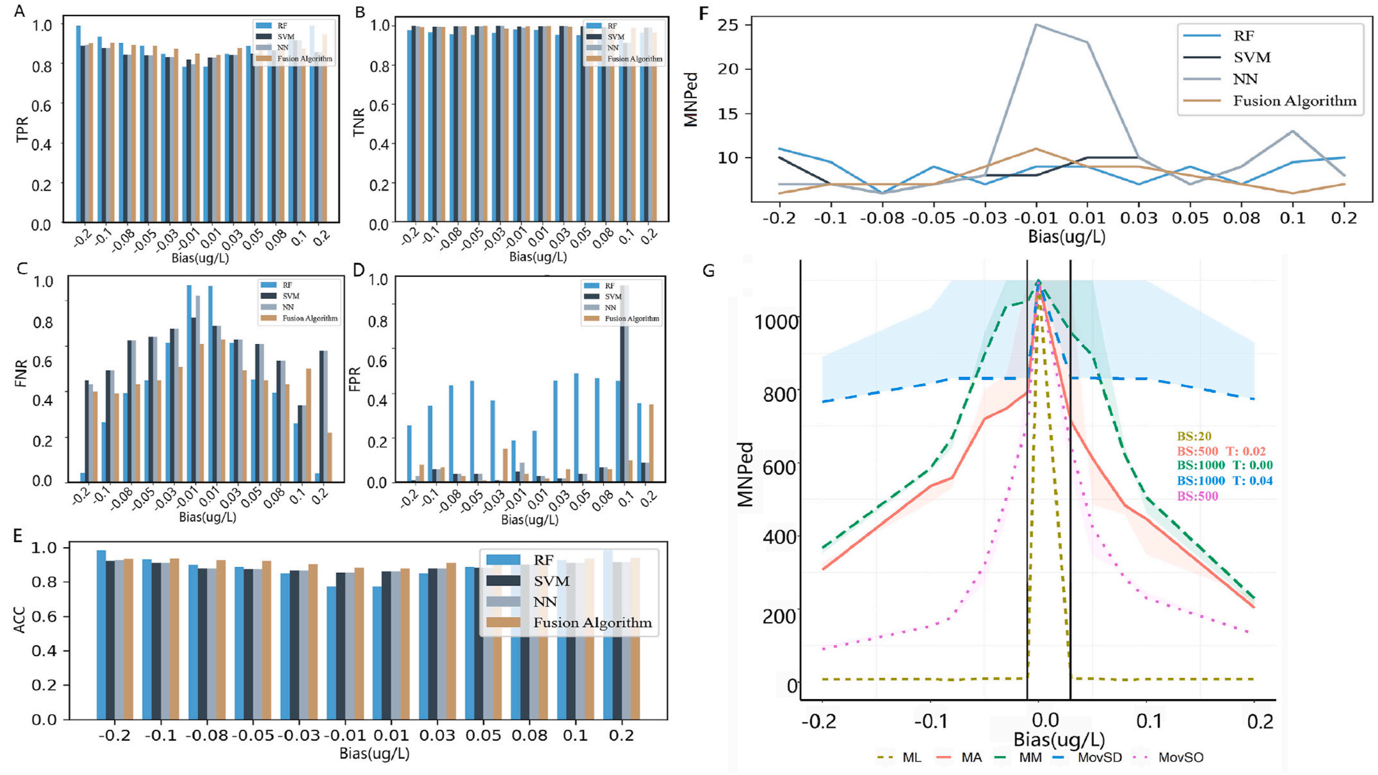

Fig. 3. The performance of ML algorithms.

A~E. TPR, TNR, FNR, FPR and ACC results of four ML algorithms for different biases.

F. MNPed of four ML algorithms for different biases.<G. MNPed of ML fusion model and four PBRTQCs for different biases.

ACC: accuracy; TPR: true positive rate: FPR: false positive rate; TNR: true negative rate; FNR: false negative rate.

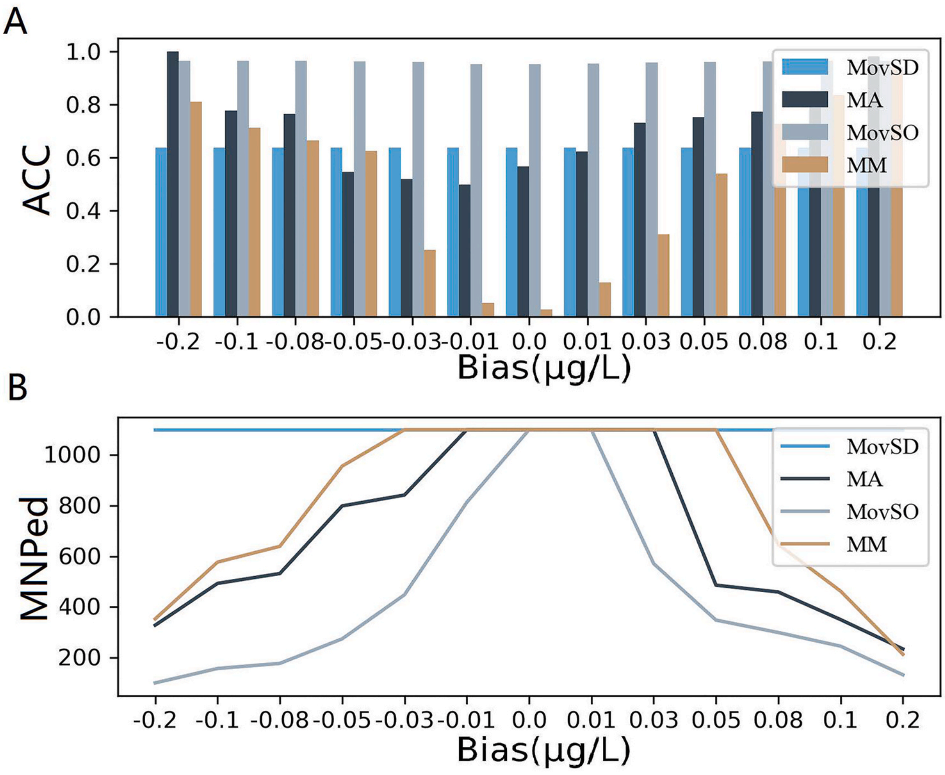

Fig. 4. The performance of four PBRTQCs.

A. ACC of 4 PBRTQCs for different biases.

B. MNPed of 4 PBRTQCs for different biases.

ACC represented accuracy. MNPed represented the median of NPed in test dataset. NPed was statistical basic unit, representing the number of patient samples from the inception of the bias until error detection.

patient results affected in real laboratory setting. And The MNPed of fusion algorithm were able to be stable below 10 for all biases, in Fig. 3G. The results of model training only for 0.15 μg/L bias were shown, due to space limitation in this paper. The 5-fold validation results of single ML model were in supplementary materials.

3.3. Comparison with PBRTQC

As a comparative method, the performance of four PBRTQCs was evaluated and the optimal PBRTQC method was selected. The overall performance order from high to low was MovSO, MA, MM and MovSD. MovSO was superior to the other three PBRTQCs in aspect of accuracy and efficiency and was more sensitive to the change of block size. Our results showed MovSO had the highest accuracy for all biases. The average accuracy of MovSO, MA, MM and MovSD for all biases were 0.96, 0.64, 0.71 and 0.51 respectively, Fig. 4A.

Delayed alarms showed the efficiency of PBRTQC for error detection. The delay of an alarm leaded to the increase of affected patient samples. Fig. 4B showed the delayed alarms of the four PBRTQCs for different biases. The straight line on the top parallel to the X axis indicated that the bias could not be detected. If a vertical line was made from both ends of the straight line on the top, we found that the number of delayed alarms each bias value in order was MovSD > MM > MA > MovSO.

Fig. 5A–D were four box diagrams, reflecting the NPed distribution of different block sizes of the four PBRTQCs when introducing 0.1 μg/L bias. It can be roughly seen that for the number of delayed alarms with different block sizes, the NPed distribution of the MovSO method was concentrated below 600, which had better performance than the other three PBRTQCs. The block sizes of MovSO and MA were only a half of that of MM and MovSD. The "neat" type of data transformation was selected for 4 PBRTQCs, the details listed in Table 2. When three different sizes of biases were introduced, MNPed illustrated the similar change trend for the four PBRTQCs, in Fig. 5E–G.

Our fusion model compared with selected optimal MovSO, MNPed results at different biases were shown in Fig. 3G, TPR, TNR, FPR, FNR, ACC, ANPed and MNPed at 0.1 μg/L biases in both directions in Table 3. The MNPed of our model was less than 12 for all biases, and ACC surpassed MovSO nearly 100 times. As shown in Fig. 3D and E, except for 0.01 μg/L bias, ACC results of our model was more than 0.9, FPR was apparently lower than MovSO, only was 0.2% and 0.1%. Even though ACC for MovSO at ±0.1 μg/L biases showed better than our model, one its MNPed was significantly higher than our model, the other TNR for MovSO was far inferior to our model roughly equal to 1. The transient better performance for MovSO may be related to our experimental design, because expected goal for searching excellent protocol was defined in advance.

4. Discussion

The incidence of prostate cancer in China has shown a significant increase in recent years. The higher proportion of patients with advanced prostate cancer in China than in the United States and other developed countries seems to be due to the lack of an early diagnosis [9]. PSA as an important indicator for screening, for monitoring and for risk group classification of prostate cancer is better than either DRE or transrectal ultrasound (TRUS). PSA is also a continuous parameter, with higher levels indicating a greater likelihood of prostate cancer (PCa), precluding an optimal PSA threshold for detecting nonpalpable but with clinically significant PCa (csPCa) [10]. The treatment monitoring goal for PSA was strict as precisely to 0.1 μg/L alteration [11]. PSA is often associated with poor prognosis or have an increased risk of metastases [12] and death [13]. The clinical significance of PSA pushes forward detection and monitoring for critical small shift from analytical source. It has been reported that the probability of errors in analytical procedure accounts for 15% of all errors, so, increasing the detection rate of small shift can reduce the rate of clinical error reporting and wrong medical decision making [5,14]. However, the accuracy of current QC method is inadequate for meeting clinical needs, especially for tPSA analyte of this kind. We try to set up a newly QC based on ML technique, improving the accuracy and efficiency of detection for small shift in real settings.

Because even a very small fluctuation from analytical sources would affect the interpretation of testing results, thus leading to improper clinical decision-making and treatment, it is necessary to select appropriate QC methods to accurately identify small analytical shift. As Fig. 2A shown, on one hand, the size of bias which needs to be detected is far less than the mode of data distribution of tPSA testing results, on the other hand, data intersection or overlapping occurs obviously between biased data and unbiased data, therefore it is difficult to separate them by traditional PBRTQC. That MLQC is superior to PBRTQC is that the separability of data is strengthened though data pre-processing and

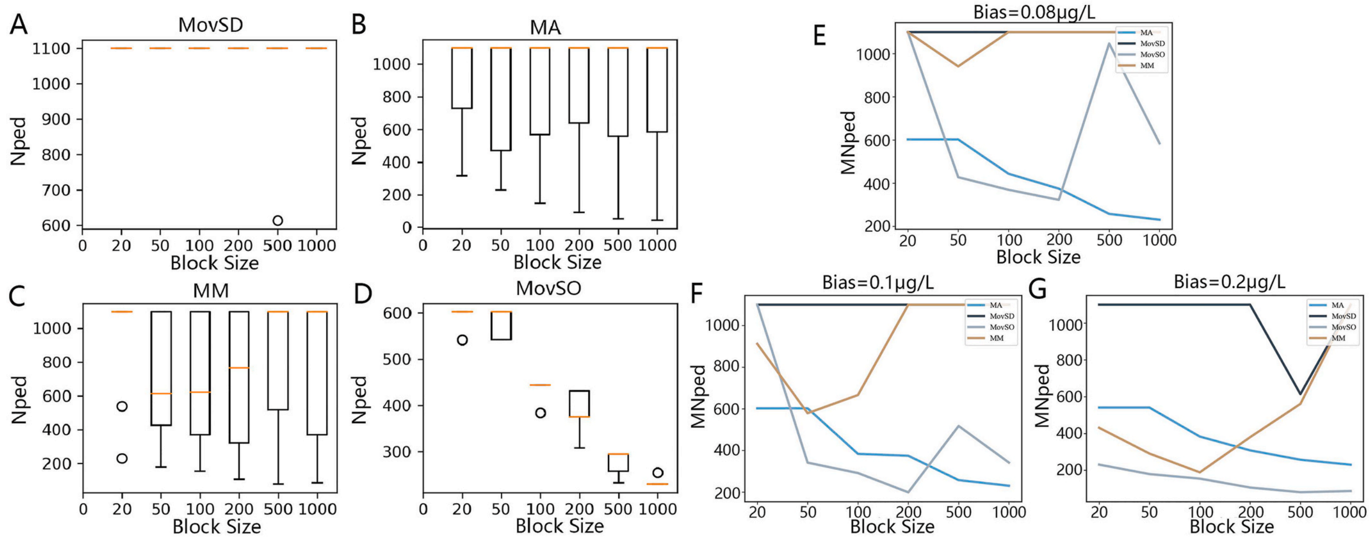

Fig. 5. The relation of NPed or MNPed to block size

A~D: Box diagrams of NPed distribution of different block sizes for four PBRTQCs.

In each diagram, small circles represented outliers, the upper and lower boundaries of the boxes represented the upper and lower quartiles, the middle horizontal lines represented the median, and the endpoints below the box represented the lower limit; E~G: MNPed of different block sizes for four PBRTQCs when introducing 0.08 μg/L, 0.1 μg/L or 0.2 μg/L bias.

Table 2
Optimal parameters for 4 PBRTQCs.

Algorithm	Truncation limit	Transformation	Block size	Control limit	CL_lower	CL_upper
MovSD	0.04	neat	1000	daily extremes	0.6623	0.7631
MA	0.02	neat	500	symmetric	1.1181	1.3714
MovSO	–	neat	500	symmetric	415.9151	447.5865
MM	0	neat	1000	all PBRTQC	0.7658	0.8835

Table 3
The performance of all algorithms at 0.1 μg/L bias both positive and negative directions for PSA test item.

Algorithm	Bias	TPR	TNR	FPR	FNR	ACC	ANPed	MNPed
MovSD	−0.1	0.725	0.792	0.208	0.275	0.726	817.5	1023.0
MA	−0.1	0.666	0.679	0.321	0.334	0.778	536.6	493.0
MovSO	−0.1	0.950	0.538	0.462	0.050	0.963	152.7	157.0
MM	−0.1	0.711	0.609	0.391	0.289	0.712	585.2	577.0
ML	−0.1	0.934	0.998	0.002	0.065	0.927	8.9	9.5
MovSD	0.1	0.752	0.659	0.341	0.243	0.758	831.0	1100.0
MA	0.1	0.747	0.407	0.593	0.253	0.832	446.2	350.0
MovSO	0.1	0.976	0.527	0.473	0.024	0.964	230.3	245.0
MM	0.1	0.835	0.698	0.302	0.165	0.835	506.1	462.5
ML	0.1	0.933	0.999	0.001	0.066	0.930	8.7	9.5

algorithm optimization.

Data filtering is the core of data pre-processing. For PBRTQC, the outlying value is removed by way of quantile truncation limits. This method is easy to remove some of hidden valuable data as the same time. Otherwise, for MLQC, when all data is mapped to high-dimensional space, the outlying value can be identified by the density and distance among data, in this way, the effective data is kept to the greatest extent. Then, data partitioning is the next step of data pre-processing. For PBRTQC, the data consist of a block size is regarded as a value; while, for ML QC, the data in one block size is regarded as different data dimensions, then cross correlation among isolated data is dig up and retained to the maximum extent. In Fig. 2B and C, each point represents a new ML sample with block sizes of 5 and 10. The previously inseparable samples have greater separability than the 4 PBRTQCs at the same block sizes. Fig. 2B and C is limited by the data visualization. The actual separability is stronger than that in Fig. 2B and C. To sum up, data pre-processing causes the loss of data information for PBRTQC. By way of enlarged block size, the information lost is made up for in certain degree. The experimental results showed that the block size of PBRTQC was nearly 100 times that of MLQC, but its improvement of error detection was limited, and the accuracy still far lagged that of MLQC.

In aspect of algorithm, MLQC, different from PBRTQC judged by control limits which is prone to false positive or false negative, recombines the isolated data in a block size, not only expanding the sample dimensions, but also adding serialization features, finally transforming a quality control problem into a multiple feature classification problem in machine learning. Our results showed that PBRTQC performed divergence for error detection to different biases. Take an example, as to MM, with the average value of absolute bias decreased by 0.01 μg/L, the accuracy was reduced by 20.8%. But the accuracy of SVM, RF and NN can reach up to 0.8, and the results are relatively stable for all biases.

As usual, due to highly complexity of clinical data in real world, single ML model is limited. For examples, RF algorithm is prone to over fitting occurrence, SVM is not good at solving multiple classification problems, and NN lacks certain decision-making ability. In this study, the weight of single a model was determined through information entropy. Then different models were combined in series according to the weight given, a multi-model fusion algorithm was produced. The fusion algorithm not only includes the mapping ability of SVM, the decision-making ability of RF and the nonlinear perceptual capabilities of NN, but also effectively removes the influence of residual errors from single model and retains the effective information of the model to the greatest extent. By the adjustment of the output of single algorithm, fusion algorithm improves the accuracy, robustness, and generalization overall. Compared with SVM, RF and NN, the accuracy of the fusion model was improved by 8.7%, 9.6% and 6.9% respectively. Our results stated that the accuracy of the fusion algorithm was 20% higher than that of PBRTQC, and the false positive rate was less than 0.002 for tPSA at 0.1 μg/L bias.

5. Limitations

This study adopts patient data to build up an innovative AI-based QC. The influence of patient data dispersion is not ignored, mainly deriving from two aspects: one is the patient population related variation, the other is the analytical variation. In the future, we will explore the method for reducing the impact of data dispersion caused by population related variation by machine learning, further improving the accuracy and generation of ML QC.

6. Conclusion

The fusion model shows outstanding performance and reduces incorrect and omitting error detection of QC, adaptable for real settings.

Research funding

This work was supported by Excellence Project of key clinical specialty in Beijing; and the 1351 Talent Training Plan [grant numbers CYMY-2017-01].

Declaration of competing interest

The authors declare that there are no conflict of interests.

Acknowledgements

The author thanks Mr. Chen Chao of Inner Mongolia Wesure Data Technology Co., Ltd for administrative management and hardware support for our experiment.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105866.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Ca - Cancer J. Clin. 71 (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] R.S. Zheng, K.X. Sun, S.W. Zhang, H.M. Zeng, X.N. Zou, R. Chen, X.Y. Gu, W. W. Wei, J. He, Report of cancer epidemiology in China, 2015, Zhonghua Zhongliu Zazhi 41 (2019) 19–28.

[3] D.Q. Sun, M.M. Cao, H. Li, S.Y. He, L. Lei, J. Peng, J. Li, W.Q. Chen, Quality assessment of global prostate cancer screening guidelines, Zhonghua Liuxingbingxue Zazhi 42 (2021) 227–233, https://doi.org/10.3760/cma.j.cn112338-20200806-01033.

[4] J. Liu, C.H. Tan, T. Badrick, T.P. Loh, Moving sum of number of positive patient result as a quality control tool, Clin. Chem. Lab. Med. 55 (2017), https://doi.org/10.1515/cclm-2016-0950, 1709-1704.

[5] J.O. Westgard, Internal quality control: planning and implementation strategies, Ann. Clin. Biochem. 40 (2016) 593–611, https://doi.org/10.1258/000456303770367199.

[6] A. Bietenbeck, M.A. Cervinski, A. Katayev, T.P. Loh, H.H. van Rossum, T. Badrick, Understanding patient-based real-time quality control using simulation modeling, Clin. Chem. 66 (2020) 1072–1083, https://doi.org/10.1093/clinchem/hvaa094.

[7] X. Duan, B. Wang, J. Zhu, W. Shao, H. Wang, J. Shen, W. Wu, W. Jiang, K.L. Yiu, B. Pan, W. Guo, Assessment of patient-based real-time quality control algorithm performance on different types of analytical error, Clin. Chim. Acta 511 (2020) 329–335, https://doi.org/10.1016/j.cca.2020.10.006.

[8] M.A. Cervinski, F.A. Polito, D. Ng, Optimization of a moving averages program using a simulated annealing algorithm: the goal is to monitor the process not the patients, Clin. Chem. 62 (2016) 1361–1371, https://doi.org/10.1373/clinchem.2016.257055.

[9] X. Li, X.Y. Zeng, Advances in epidemiology of prostate cancer in China, Cancer. Res. Prev. Treat. 48 (2021) 98–102, https://doi.org/10.3971/j.issn.1000-8578.2021.20.0370.

[10] N. Mottet, R.C.N. van den Bergh, E. Briers, T. Van den Broeck, M.G. Cumberbatch, M. De Santis, S. Fanti, N. Fossati, G. Gandaglia, S. Gillessen, N. Grivas, J. Grummet, A.M. Henry, T.H. van der Kwast, T.B. Lam, M. Lardas, M. Liew, M.D. Mason, L. Moris, D.E. Oprea-Lager, H.G. van der Poel, O. Rouvière, I.G. Schoots, D. Tilki, T. Wiegel, P.-P.M. Willemse, P. Cornford, EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. Part 1: screening, diagnosis, and local treatment with curative intent, Eur. Urol. 79 (2021) 243–262, https://doi.org/10.1016/j.eururo.2020.09.042.

[11] G. Ploussard, F. Staerman, J. Pierrevelcin, R. Saad, J.-B. Beauval, M. Roupret, F. Audenet, M. Peyromaure, N. Barry Delongchamps, S. Vincendeau, T. Fardoun, J. Rigaud, A. Villers, C. Bastide, M. Soulie, L. Salomon, Predictive factors of oncologic outcomes in patients who do not achieve undetectable prostate specific antigen after radical prostatectomy, J. Urol. 190 (2013) 1750–1756, https://doi.org/10.1016/j.juro.2013.04.073.

[12] D.E. Spratt, D.L.Y. Dai, R.B. Den, P. Troncoso, K. Yousefi, A.E. Ross, E.M. Schaeffer, Z. Haddad, E. Davicioni, R. Mehra, T.M. Morgan, W. Rayford, F. Abdollah, E. Trabulsi, M. Achim, E.L.N. Tapia, M. Guerrero, R.J. Karnes, A.P. Dicker, M. A. Hurwitz, P.L. Nguyen, F.F.Y. Feng, S.J. Freedland, J.W. Davis, Performance of a prostate cancer genomic classifier in predicting metastasis in men with prostate-specific antigen persistence postprostatectomy, Eur. Urol. 74 (2018) 107–114, https://doi.org/10.1016/j.eururo.2017.11.024.

[13] F. Preisser, F.K.H. Chun, R.S. Pompe, A. Heinze, G. Salomon, M. Graefen, H. Huland, D. Tilki, Persistent prostate-specific antigen after radical prostatectomy and its impact on oncologic outcomes, Eur. Urol. 76 (2019) 106–114, https://doi.org/10.1016/j.eururo.2019.01.048.

[14] M. Plebani, P. Carraro, Errors in a stat laboratory: types and frequencies 10 Years later, Clin. Chem. 53 (2007) 1338–1342, https://doi.org/10.1373/clinchem.2007.088344.