Clinica Chimica Acta 495 (2019) 233–238

Contents lists available at ScienceDirect

Clinica Chimica Acta

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

Quality control optimization part II: A method to optimize the accuracy of laboratory quality control

Robert L. Schmidt, Lauren N. Pearson*

Department of Pathology, University of Utah, Salt Lake City, UT, United States of America

ARTICLE INFO

Keywords:
Quality control
Error
Optimization
Accuracy
Statistics
Sensitivity
Specificity

ABSTRACT

Background: Quality control (QC) can be viewed as a diagnostic test that is used to determine whether an assay is in statistical control. Using this framework, QC performance can be evaluated using familiar metrics associated with diagnostic tests. QC plan parameters can be adjusted to optimize performance metrics.
Methods: We developed a simple dichotomous model based on classification of assay errors. Errors are classified as important or unimportant based on a critical shift size, defined as Sc. Using this scheme, we show how QC policies can be analyzed using common accuracy metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We conducted computer experiments to determine the QC plan that optimizes QC accuracy under a wide range of scenarios.
Results: In general, traditional QC plans (based on2 or 3 standard deviation limits) are approximately 90% as accurate as optimized QC limits in the scenarios that were tested. There are special circumstances when traditional QC plans do not perform well.
Conclusion: QC performance can be optimized for specific contexts.

1. Introduction

Quality control (QC) is a critical activity in clinical laboratories because QC insures the reliability of test results. Laboratories spend significant amounts of time and money on quality control. Quality control activities account for significant portion of a laboratory's budget. Given this effort, it is important to ensure that QC procedures are effective.

QC plans are generally designed according to traditional rules of thumb. For example, two standard deviations or three standard deviations are commonly used limits for QC monitoring. Results that exceed these limits are considered unusual and provide evidence that the measurement process has changed and that results may be unreliable. Such rules have proved useful in practice but it is unclear how to evaluate such rules. Although CLSI has provided guidance for developing QC strategies [1], it is unclear whether traditional rules are optimal with respect to some objective such as cost or accuracy.

Traditionally, QC monitoring plans have been evaluated using power curves. Given a shift in the mean of the process, power curves provide the probability of detecting the shift. Based on power curve analysis, two types of errors can occur. First, a rule failure can occur when no shift has occurred (false rejection). Second, the QC monitoring system can fail to generate a rule failure when a shift has occurred (false negative). This scheme treats all shifts equally. Both large and small shifts are considered errors. In reality, the magnitude of the shift is important. Large errors are more important than small errors and are also easier to detect. Many small shifts are inconsequential and can be ignored. For example, a shift of three standard deviations (SD) is likely to be important whereas a shift of 0.2 SD might be insignificant. Time and money are consumed in the troubleshooting process so it is important to respond only to important events. Also, the troubleshooting is most likely to be productive if it is applied to significant events.

In a previous study, we developed a simple method for classifying upsets (i.e. shifts) as important or unimportant. In that paper, we described how a QC monitoring plan is similar to a diagnostic test. Like a diagnostic test, a QC monitoring plan provides a result that indicates whether an important event has occurred. Using that framework, we showed how QC performance can be evaluated using performance measures that are familiar to laboratory personnel (sensitivity, specificity, positive predictive value, negative predictive value, and accuracy). In this study, we extend that work to show how the framework can be used to optimize QC plans. The objective of this study is to calculate QC limits that optimize the accuracy of QC monitoring and compare the performance of an optimal monitoring plan to traditional QC monitoring plans. To do so, we extend a simple model which appears in a companion paper [2].

* Corresponding author at: Department of Pathology, 15 N Medical Drive East, Suite 1100, University of Utah, Salt Lake City, UT 84112, United States of America.
E-mail address: lauren.pearson@aruplab.com (L.N. Pearson).

https://doi.org/10.1016/j.cca.2019.04.054
Received 20 December 2018; Received in revised form 22 February 2019; Accepted 5 April 2019
Available online 17 April 2019

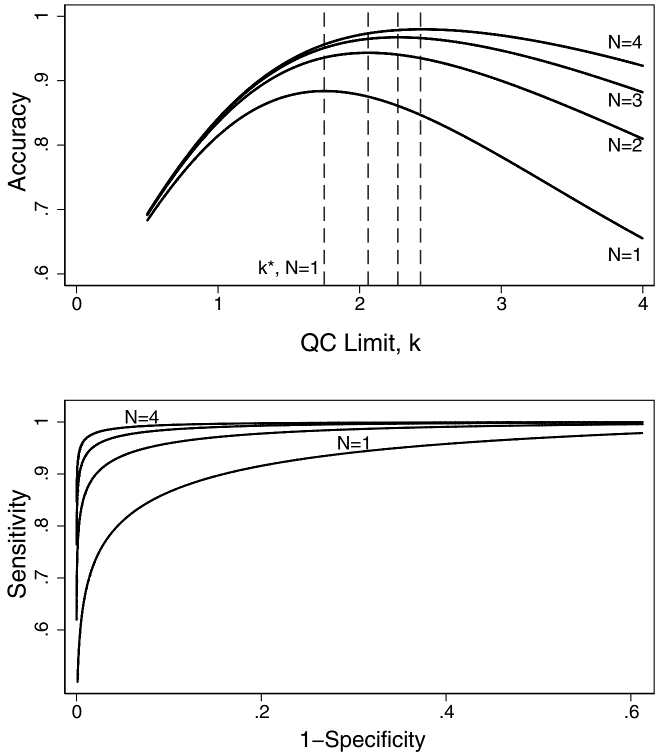

Fig. 1. The impact of the quality control (QC) limit on accuracy of QC. The Fig. shows data for a critical shift size Sc = 1.5σ and event probability $p = .01$. N is the number of QC repeats at one concentration level. K is the QC limit expressed as multiples of the standard deviation. Each value of k is a scenario that represents a different QC plan. The dashed lines show the optimum value of k (k*) for each number of repeats, N. For example, k* for $N = 1$ is indicated in the graph. The lower panel shows an ROC curve for the same data.

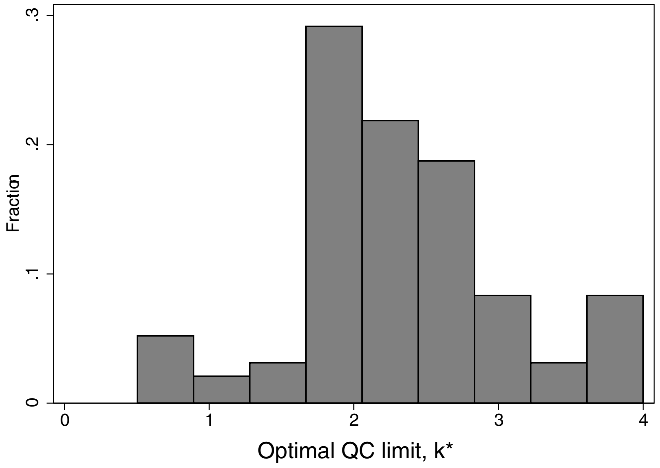

Fig. 2. Distribution of optimal quality control limit. The fig. Shows the optimal QC limit for 96 different scenarios which varied with respect to the critical shift value (Sc), the number of repeats (N) and the event rate (p). The QC limits were adjusted to maximize accuracy.

2. Methods

We assume that upsets occur randomly and follow a Bernoulli distribution with parameter p. Because we have no prior knowledge of the distribution of shift sizes, we assume that shifts have a uniform distribution ranging from 0 to 5 SDs: $g(S|E)\sim U(0,5)$ (Our previous work showed that our model was relatively insensitive to assumptions regarding the distribution of shifts). We assume N ∊ {1,2,3,4}, Sc ∊ {0.5, 1, 1.5, 2, 2.5, 3}, and p ∊ {0.001, 0.01, 0.1, 1}. The combinations for N, Sc and p, formed 96 scenarios (4 levels for N, 6 levels for Sc, 4 levels for p).

For each scenario, we used Monte Carlo simulation to create 1000 sub scenarios in which k randomly varied between 0.5 and 4. We calculated performance statistics (sensitivity, specificity, positive predictive value, negative predictive value, and accuracy for each sub scenario. The optimum was defined as the value of k, which maximized the accuracy for each scenario (Fig. 1). Given a value of Sc, we compared the performance of various QC plans (k = 2, k = 3, and k*). We selected k = 2 because a 2σ limit is most widely used in clinical laboratories. We selected k = 3 because this limit is incorporated in Westgard rules and is the traditional limit for QC outside of clinical laboratory medicine [3]. We then determined the relative accuracy that would be obtained if one used a traditional QC limit (k = 2 or k = 3) rather than an optimized QC limit (i.e., k = k*). We also determined the relative performance for sensitivity, specificity, PPV and NPV. We used regression analysis to obtain an equation to predict the optimal QC limit given a scenario (Sc, p, N).

Calculations were performed using Microsoft Excel. Simulations were performed using @Risk 7.0 (Palisade Corp.) which is an Excel add-in. Regression analysis was performed using Stata 14 (Stata Corp.). Graphics were also produced by Stata 14.

3. Results

Optimal values of k were determined for 96 scenarios (Fig. 2, Supplementary Table 1). The optimal value ranged from 0.5 to 4.0 (median 2.14: interquartile range: 1.9–2.8, Fig. 2). We found that the optimal values of k could be described by the following relations:

$$k^* = \begin{cases} p < 1\text{: } 1.78 - 0.39S_c + 0.14N + 0.12S_c^2 + 0.16S_cN\ (R_{adj}^2 = 0.98) \\ p = 1\text{: } -0.98 + 1.20S_c + 0.47N\ (R_{adj}^2 = 0.93) \end{cases}$$

We investigated how the optimal QC limit changed relative to the critical shift size (Fig. 3). The patterns were almost identical when $p \leq .1$. In that case, the ratio k*/Sc was about 4 when Sc was 0.5 and decreased to about 1 when Sc increased to three. The pattern was strikingly different when $p = 1.0$.

Optimizing the QC limits resulted in a modest improvement in accuracy relative to traditional QC limits of 2 and 3 standard deviations (Fig. 4). The accuracy of the 2 SD limit was generally at least 95% of the accuracy of the optimal control limit. The accuracy of the 3 SD limit was generally at least 85% of the accuracy of the optimal control limit.

Optimizing QC limits for accuracy affected the other non-optimized performance measures. Optimization had mixed effects on PPV, NPV, sensitivity and specificity relative to the performance with 2 SD limits (Fig. 5). In most cases, performance was unchanged; however, there was considerable variance. For example, the ratio PPV(k = 2)/PPV (k = K*) ranged from 0.02 to 1.86 (median 0.87). We found that NPV, sensitivity, and specificity associated with the 3 SD limit were generally lower than the values obtained using the optimized limit (Fig. 6). In contrast, the PPV associated with 3 SD limit ranged from 0.3 to 18.7 relative to the PPV associated with the optimized limit (median 1.54, interquartile range: 1.1–4.9).

4. Discussion

We developed a simple method for optimizing QC limits based on a simple classification of errors into two categories: important and

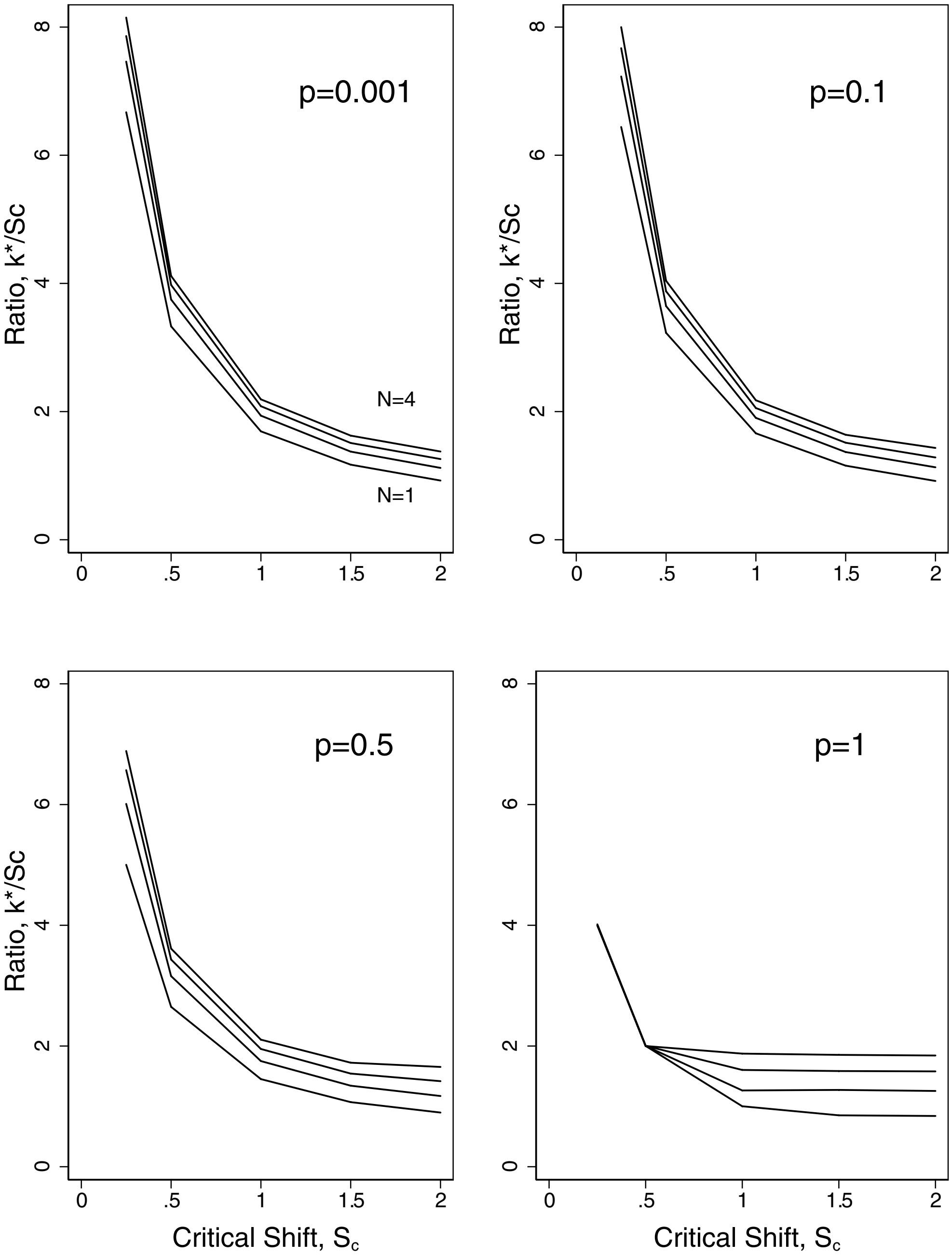

Fig. 3. Ratio of optimal quality control (QC) limit, k*, to critical shift size, Sc, as a function of critical shift size. The fig. Shows how the optimal QC limit is placed relative to the critical shift size. N is the number of repeats and p is the event rate.

unimportant. By doing so, QC can be regarded as a diagnostic test that can be evaluated using common performance measures such as sensitivity, specificity, PPV, NPV and accuracy. Using accuracy as a goal, we calculated the optimum QC limits and investigated the impact of optimization on the common performance measures.

As expected, changing QC limits to optimize accuracy improved the accuracy of QC relative to the traditional QC limits (2 SD, 3 SD). We compared performance over a wide range of scenarios and found that optimization of QC limits provided modest gains in accuracy over traditional QC limits. The accuracy achieved with optimized limits was about 5% better than the accuracy obtained with 2 SD limits and about 10% better than the accuracy obtained with 3 SD limits.

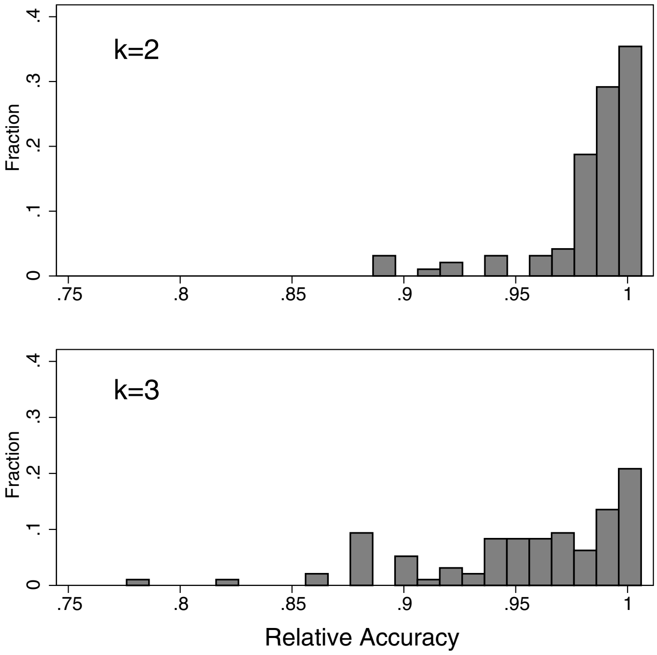

Fig. 4. Relative Accuracy. The fig. Shows ratio of the accuracy using traditional quality control limits (k = 2, k = 3) relative to the accuracy obtained using optimized control limits (k = k*).

We were somewhat surprised that optimization did not lead to greater improvement in accuracy. Optimization definitely improves accuracy but the gains were relatively modest – particularly relative to 2 SD limits. Our results suggest that 2 SD limits are generally more accurate than 3 SD limits if one were to choose between traditional QC plans. Although the optimal QC limit varied we found that the optimal QC limits were most often in the range of 2 SD to 3 SD (Fig. 2). Although we were surprised that optimization provided only modest gains in accuracy, the fact that traditional limits are roughly optimal should not be surprising given the long history of QC. QC limits would have been adjusted if they were not useful. Although the gains were generally modest, there were isolated cases with significant differences in performance.

Our model is based on a concept similar to Westgard's critical shift size. In the Westgard approach a medically important shift is defined as one which would lead to 5% of the results exceeding the total allowable error. Our approach is similar but more flexible. We believe that the definition of an important shift should depend on the decision context. QC can be used for different purposes. For example, it can be used to insure compliance or to drive operations improvement. In a medical context, the shift size might be adjusted according to medical risk. Similarly, the critical shift size used in a operations context (e.g., to drive process improvement) would most likely differ from the shift size used in a medical context. Although our method can be applied to laboratory test results, our method is quite general and can be applied to any control chart.

We chose accuracy as the criterion for optimization and we varied the QC limit to achieve the maximum accuracy. Changing the QC limits affects other performance measures (sensitivity, specificity, PPV, NPV). We found that optimizing QC limits for accuracy usually had modest effects on the other performance measures. One notable exception was the impact on the PPV when using a 3 SD limit. In general, the PPV obtained using a 3 SD limit is significantly higher than the PPV obtained with optimized limits. This result is not surprising. A 3 SD limit is designed to have a high PPV. Optimizing for accuracy usually results in QC limits that are lower than 3 SD and, as result, the PPV of the optimized QC limit is lower. The fact that optimization of accuracy sometimes led to decreases in other performance measures is expected. Optimizing any single performance criterion will have side effects and cause changes in other performance measures.

Accuracy is only one of many choices for optimization. We selected accuracy because it requires relatively few assumptions. However, our model provides a basis optimizing with respect to many other criteria. For example, one could minimize costs by assuming a relative cost for false negatives and false positives. Or, one could optimize multiple objectives such as cost, accuracy or predictive values (multi-objective optimization) or optimize a parameters subject to constraints. For example, one could optimize accuracy subject to a constraint that the PPV and NPV exceed certain targets. One might also optimize metrics such as number of QC events to detect an out-of-control event or the number of events before a false rejection.

Our model differs from other approaches because we study average performance over many events. In traditional QC, one assumes an event (e.g., a shift of size S) and then uses a power curve to determine the probability of error detection and false rejection. These parameters enable one to calculate metrics recommended by CLSI such as the expected number of QC events between false rejections or the expected number of QC events to detect an out-of-control condition [1]. The CLSI approach performs this analysis with respect to a single event. In contrast, our model optimizes global performance with respect to an accuracy objective. Thus, it is difficult to use the CLSI criteria to evaluate the performance of our model. Our model could be modified to optimize the CLSI metrics. This may be an opportunity for future research.

Our model is limited because it depends on assumptions regarding out-of-control behavior and little is known about the distribution of upsets. We employed a very flexible model that is based on two components: 1) the frequency of upset events 2) the magnitude of those events (i.e., shift sizes). In a companion article, we investigated the impact of the distribution of magnitude of shifts. We compared a uniform distribution (all shift sizes equally likely), an exponential distribution (small shift sizes more likely). and a triangular distribution (shift sizes tend to be clustered). We found that the distribution of shifts had a relatively small impact on QC performance. For that reason, we used a single distribution (uniform distribution) for shifts in this study. We considered a wide range of event rates ($0.001 < p$ 1) and found that the optimal limit was similar for a wide range of scenarios. ($p \leq .1$). Thus, although our method required us to make assumptions regarding upset events, our results do not appear to be sensitive to these assumptions.

Our model also has several strengths. The most significant strength is its simplicity. Our method only requires operators to define a boundary between important and unimportant events. We believe most operators would be able to provide this. Other researchers have

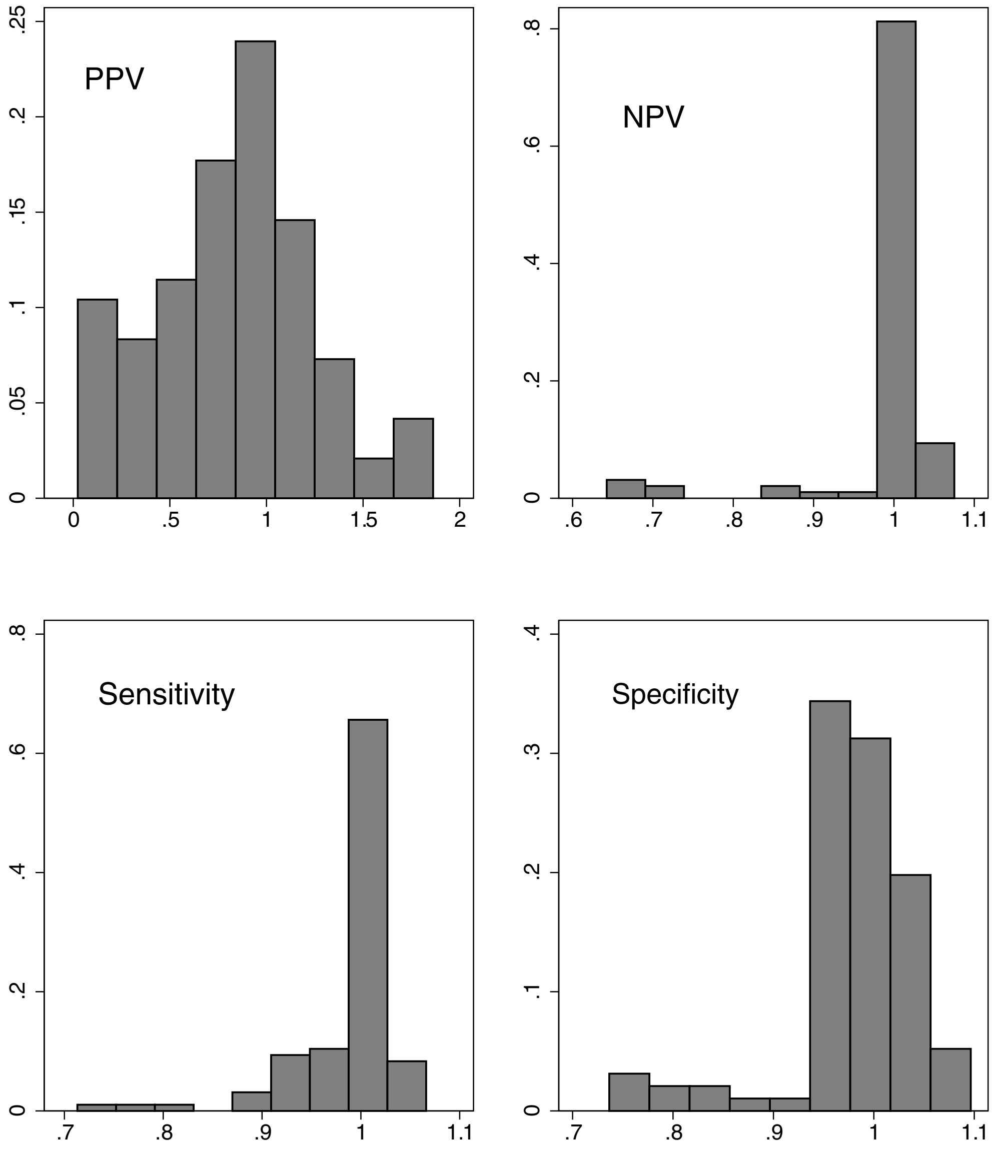

Fig. 5. Impact of optimization accuracy on other performance measures. The fig. Shows the ratio of the performance measure using a traditional quality control limit (k = 2) to the performance using an optimized limit (k = k*). Optimization was performed by maximizing accuracy by adjusting the control limit. These performance measures were not optimized but were affected by changing the quality control limit from k = 2 to k = k*. (k* is the QC limit that optimizes accuracy). A ratio less than one indicates that optimization improved the performance parameter.

developed models to optimize QC; however, in general, these require complicated assumptions about cost functions. Our model is based on a simple dichotomous classification which enables one to employ common performance measures associated with diagnostic tests that are familiar to laboratory personnel. Our model could be extended to optimize costs, or other criteria, but we think it is an advantage to use performance measures that are familiar to laboratory personnel and require relatively few assumptions.

In summary, we developed a simple method to optimize the accuracy of laboratory quality control. Our results suggest that the method would lead to modest improvements in accuracy in the majority of scenarios but could provide significant benefits in isolated scenarios.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2019.04.054.

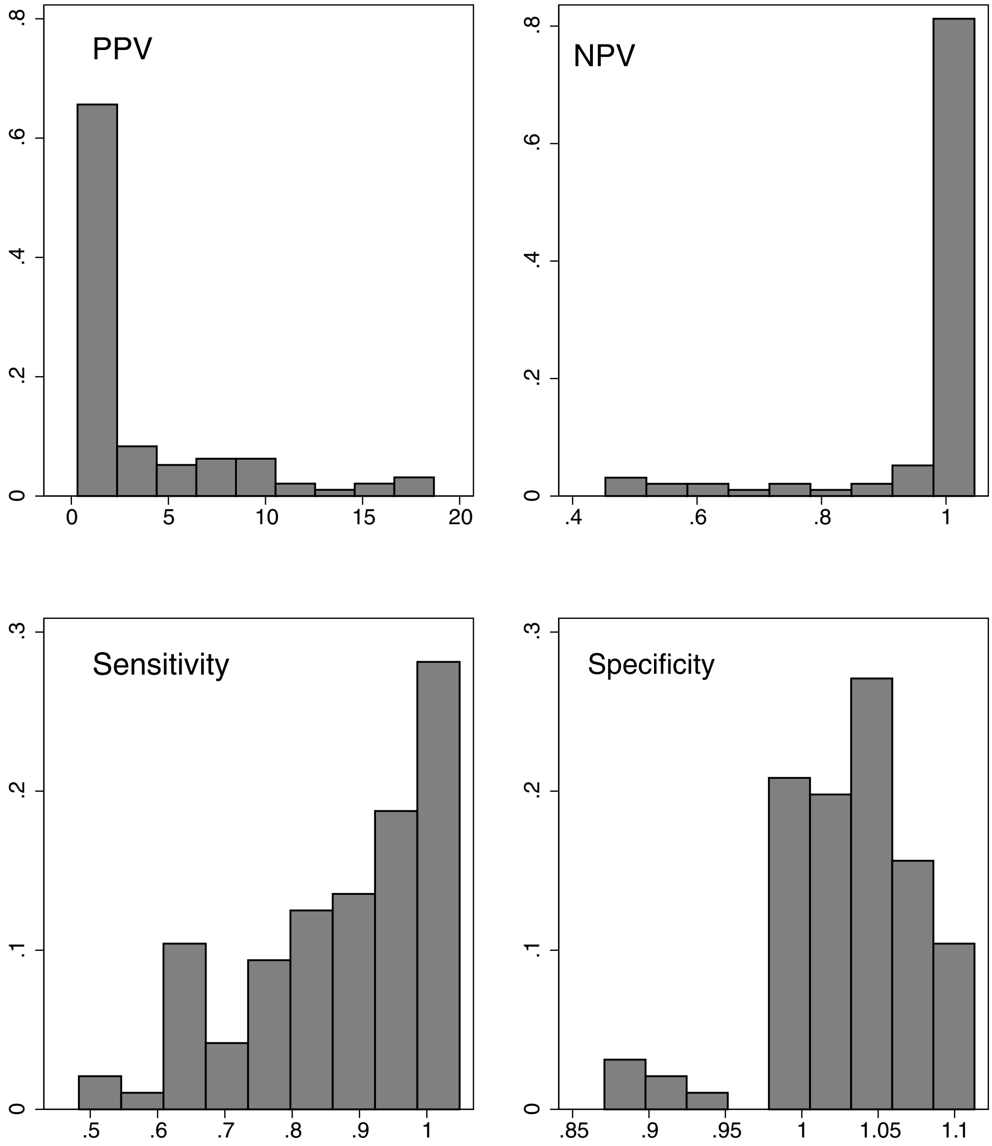

Fig. 6. Impact of optimization of accuracy on other performance measures. The fig. Shows the ratio of the performance measure using a traditional quality control limit (k = 3) to the performance using an optimized limit (k = k*). Optimization was performed by maximizing accuracy by adjusting the control limit. These performance measures were not optimized but were affected by changing the quality control limit from k = 2 to k = k*. (k* is the QC limit that optimizes accuracy). A ratio less than one indicates that optimization improved the performance parameter.

References

[1] CLSI, Statistical Quality Control for Quantitative Measurement Procedures: Principles and Definitions, Clinical Laboratory Standards Institute, Wayne, PA, 2016.

[2] R.L. Schmidt, L. Pearson, Quality control optimization part I: metrics for evaluating predictive performance of quality control, Clin. Chim. Acta 495 (2019) 174–184.

[3] J.O. Westgard, T. Groth, Power functions for statistical control rules, Clin. Chem. 25 (1979) 863–869.