Clinica Chimica Acta 495 (2019) 174–184

Contents lists available at ScienceDirect

Clinica Chimica Acta

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

Quality control optimization part I: Metrics for evaluating predictive performance of quality control

Robert L. Schmidt[a,b], Lauren N. Pearson[a,*]

[a] The Department of Pathology, University of Utah, Salt Lake City, UT, United States of America
[b] ARUP Laboratories, Salt Lake City, UT, United States of America

ARTICLE INFO

Keywords:
Quality control
Statistics
Error
Positive predictive value
Negative predictive value

ABSTRACT

Background: Quality control (QC) policies are usually designed using power curves. This type of analysis reasons from a cause (a shift in the assay results) to an effect (a signal from the QC monitoring process). End users face a different problem: they must reason from an effect (QC signal) to a cause. It would be helpful to have metrics that evaluated QC policies from an end-user perspective.

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 explore the impact of design choices (QC limits, number of repeats) on these performance measures in a number of different contexts.

Results: PPV varies widely (1% to 100%) depending on context. NPV also varies (40% to 100%) but is less sensitive to context than PPV. There are many contexts in which QC policies have low predictive values. In such cases, performance (PPV, NPV) can be improved by adjusting the QC limits or the number of repeats at each QC event.

Conclusion: The effectiveness of QC can be improved by considering the context in which the QC policy will be applied. Using simple assumptions, common accuracy metrics can be used to evaluate QC policy performance.

1. Introduction

Laboratories are under increasing pressure to improve performance. Quality control (QC) ensures the reliability of results and is therefore a key component of laboratory performance. Laboratories direct considerable resources to QC and assay improvement and, given the importance of QC, it would be useful to have metrics to evaluate the performance of a QC plan. Unfortunately, few metrics are available.

The performance of a QC plan is generally analyzed in terms of the number of events before false rejection and the number of events before error detection [1]. These quantities are also known as the average run length (ARL) and time to signal (TTS) [2]. Run lengths are determined by the statistical power of a QC plan. Statistical power is the probability that a QC plan will produce a signal (i.e., rule violation) when a change in the process occurs (e.g., a shift in the mean) [2–4]. QC plans with greater statistical power are considered superior. Such analyses fail to consider the magnitude of the error and all errors are considered equal. In reality, this assumption is unlikely to be true because larger errors may have more potential for harm (and may be costlier) than smaller errors. A more accurate model might place more weight on larger errors. In particular, power curve analysis only considers an event a false rejection when the QC monitoring system produces a signal and there has been no shift in the mean. This practice overstates the specificity of the QC plan because there may be inconsequential events (i.e., small shifts), which can be safely ignored. Responding to such events wastes resources. A more realistic model might classify errors into categories (e.g., important/unimportant) and use this information to evaluate the performance of a QC plan.

The design of QC plans is rarely considered from a user perspective. The typical design perspective is, “Given a shift of a given size, what is the probability of detecting the change if I use a particular QC plan?” The reasoning is from cause to effect. The end-user perspective is different. The end user is confronted with a QC result and asks, “Given this signal, what is the probability that a significant problem has occurred? Is it worth the time to troubleshoot?” Conversely, “Given no signal, what is the probability that no change has occurred?” End users need to reason from an effect (a signal from QC monitoring) to a cause.

A QC monitoring plan can be viewed as a statistical test that

* Corresponding author at: Department of Pathology, University of Utah, 15 N Medical Drive East, Suite 1100, 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.2019.04.053
Received 20 December 2018; Received in revised form 22 February 2019; Accepted 5 April 2019
Available online 08 April 2019
0009-8981/

determines whether a significant change in operations has occurred (e.g., shift in the mean) [5]. Although metrics based on power curves (ARL and TTS) provide useful information, they are based on assumptions regarding the underlying state of the process. Because QC can be viewed as a diagnostic test, it would be helpful to evaluate QC strategies using common measures of accuracy such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) because these terms are familiar to laboratory personnel. Predictive values would be particularly useful because they address the questions that end users face when presented with QC results.

Westgard and Groth showed how QC monitoring can be viewed as a diagnostic test [5]. Using a simple model, they showed how the predictive values of QC depend on the prevalence of errors. In their model, Westgard and Groth assumed values for the probability of error detection, P_{ed}, and false rejection, P_{fr}. While their model is correct, the estimates of predictive values depend on P_{ed} and P_{fr} and they did not attempt to estimate these statistics. The objective of this study is to extend the work of Westgard and Groth by studying the impact of error distributions on the predictive values. To that end, we provide a model that predicts impact of error distributions on the performance of QC plans.

2. Theoretical development

2.1. Background

Our goal is to develop a probability model that can be used to compare various QC plans. A probability model is created by describing all possible outcomes and defining events. An event is a set of outcomes. Probabilities are then assigned to the events. (Set notation is described in Appendix A. A list of variables is provided in Appendix B).

2.2. Model

We consider QC monitoring at a single level (i.e., 1 concentration) in a batch process. We wish to determine whether a shift in the mean has occurred in the current batch. (We recognize that QC is usually performed at multiple concentrations. The method we describe could be independently applied at each concentration). We define a QC event as an occasion in which N measurements ($N \geq 1$) are performed at a particular time ($t = 1, 2, \ldots$). Let $X_{i,\,t}$ designate the i^{th} measurement at time t. We assume the individual QC results, $X_{i,\,t}$, are normally distributed with mean, μ, and SD, σ. The system has a baseline mean $\mu = \mu_0$ when the system is in statistical control. Each QC event is associated with a sample average, $\overline{X}_t = \frac{1}{N}\sum_i X_{it}$.We will assume the system is monitored using a $k\sigma_{\overline{x}}$ QC limit (QC fails if $\overline{X}_t > k\sigma_{\overline{x}}$). The number k represents the control limit in terms of SD units. Typical values of k are 2 or 3 SDs. The monitoring system can be viewed as a test for out-of-control events. We will define a QC rule violation event, T, as $|\overline{X}_t| > k\sigma_{\overline{x}}$. The event T^c (T complement), $|\overline{X}_t| < k\sigma_{\overline{x}}$ occurs when there is no violation of the QC rule. A QC plan, $\pi(N,k)$, is a particular choice of QC limit, k, and number of repeat samples, N, taken at each QC event.

We are interested in the ability of various QC plans to detect shifts in the mean of the process. We assume that such shifts are sporadic. Let S designate the magnitude of the shift ($S \geq 0$) expressed as a multiple of the SD, σ. We will refer to such events as "upset" events. After such an event, the mean of the distribution is given by $\mu = \mu_0 + S\sigma$. Each QC event is associated with a random variable, $S \geq 0$. We define $E = \{S : S > 0$ as the event that a positive ($S > 0$) shift occurred (E^c is the event that a shift of size zero occurred). The magnitude of positive shifts is described by a distribution, $g(S|E)$, which could take various forms. Thus, shifts are determined by two components: 1) the probability that a shift will occur, $P(E) = p$ and 2) the size of the shift given that a shift occurred, $g(S|E)$. For example, suppose that p = .01, and the g (S|E) is a uniform distribution with range (0,5]. Given these inputs, a shift event would occur in one out of every 100 QC events. When such an event occurs, the size of the shift would be randomly selected from the interval (0,5]. Overall, this model allows for random occurring shift events with random shift sizes. This is a flexible model that can be used to study a wide range of upset patterns by varying the upset rate, p, and the pattern of shift sizes, g (S|E) (Fig. 1).

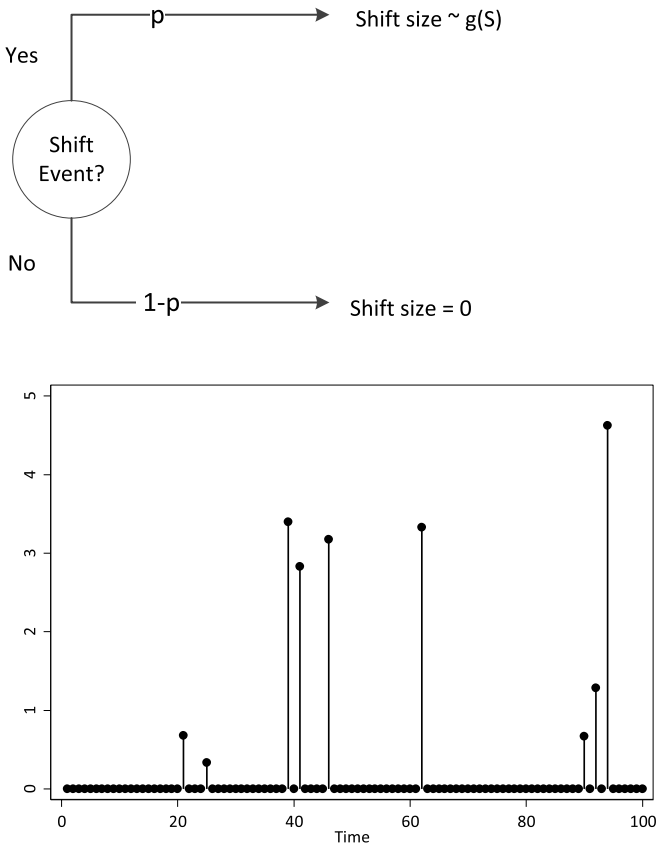

Fig. 1. Model for Shifts. At each QC event, there is a chance, p, that there will be a shift in the mean of the results. If a shift occurs, the size of the shift, S, is determined by the probability distribution, g(S). This is a flexible model in which the event rate, p, and the distribution of shift sizes can be varied by choosing different event rates and shift distributions. Upper panel: decision tree showing possible outcomes at each QC event. The lower panel shows the shift distribution from 100 QC measurement events in which the event rate is p = .1 and the shift distribution is uniform (minimum = 0, maximum = 5). The Figs. shows that random shifts with sizes between 0 and 5 occur at about 10% of the time points.

We are only interested in detecting "important" shifts because some shifts are inconsequential and can be ignored. Small shifts are unlikely to compromise the reliability of results and troubleshooting such events is unlikely to be productive. We wish to avoid the effort and expense that is associated with investigating QC failure events. We will designate a critical shift size, S_c, that separates important shifts from inconsequential shifts (Fig. 2). Shifts above S_c are considered important and shifts below S_c are considered unimportant. We wish to compute

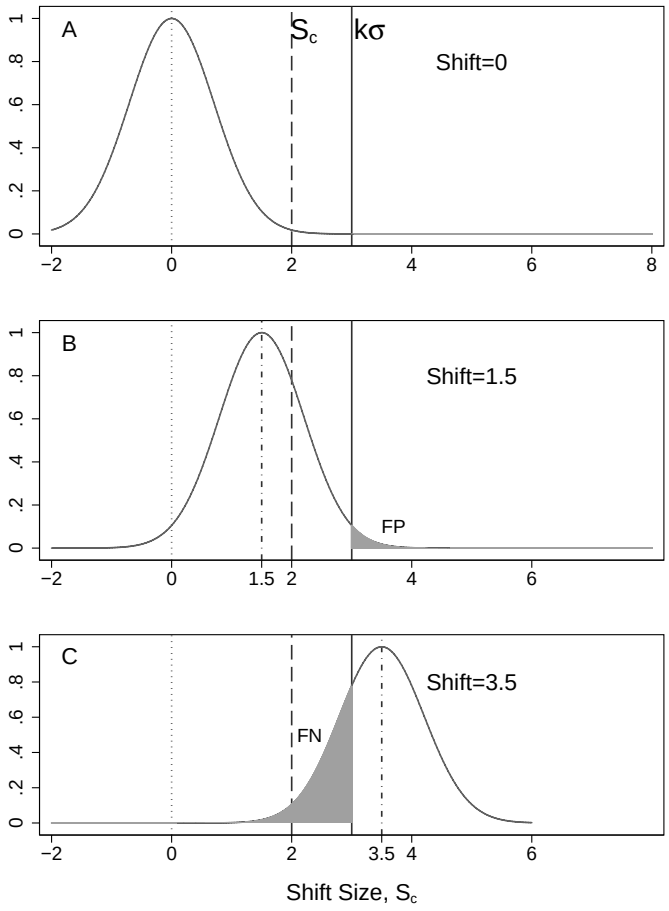

Fig. 2. Description of Model. The critical shift, S_c, is a decision point that separates important shifts from unimportant shifts. The control limit, kσ, is shown as the solid line. A result greater than the control limit causes a rule violation (run failure). The Figs. show the distribution of results for three different shifts. A false-positive result occurs when the shift (mean of the distribution) is less than critical shift but the QC result is greater than the control limit (panel B). A false-negative result occurs when the shift (mean of the distribution) is greater than critical shift but the QC result is less than the control limit (panel C).

Table 1
Scheme for evaluating accuracy of QC rules.

QC rule violation	Shift size, S	
	Important (S > Sc)	Unimportant (S < Sc)
Yes	True positive	False positive
No	False negative	True negative

Shifts occur in the mean of the assay results. The shifts are classified as important or unimportant relative to a critical shift size, Sc. A shift is considered important if it is greater than Sc. The quality control (QC) plan is designed to detect important shifts and ignore unimportant shifts. The QC result can be viewed as a diagnostic test to reveal the underlying shift. A true positive occurs when a QC rule is violated and an important shift in the mean occurs.

the probability of detecting important shifts with a given QC plan.

By categorizing shifts into two categories, a QC result can be viewed as a diagnostic test that provides information about the underlying shift. Ideally, rule violations would occur when an important shift occurred (true positive result), and no rule violation would occur when an unimportant shift occurred (true negative result). Categorizing shifts into two categories enables us to apply all the usual accuracy metrics (sensitivity, specificity, predictive values) to evaluate the performance of a QC plan (Table 1, Fig. 3). The accuracy of a QC plan means the ability of a QC plan to discriminate between important and unimportant events. The ability to discriminate between such events could be useful, and accuracy metrics could be used to evaluate the performance of a QC plan. The mathematic derivation of the accuracy metrics is provided in Appendix C.

3. Methods

The probability model was implemented in Microsoft Excel. We used the probability model to determine the impact of various QC plans on performance measures such as the sensitivity, specificity, PPV, and NPV. We investigated three distributions of shifts g(S|E): 1) uniform ($0 < S < 5$), 2) exponential ($\lambda = 1$), and 3) triangular (minimum = 0, mode = 2, maximum = 4). These distributions are presented in Fig. 4. The uniform distribution was used to model a context where shifts of any size are equally likely. (The uniform distribution is a "noninformative" distribution and is often used to model situations where there is little knowledge of the underlying behavior.) The exponential distribution was used to model a context where small shifts are more likely than large shifts. A triangular distribution (min = 0, mode = 2, max = 4) was used to model a context where shifts tend to be clustered around a mean. We selected four levels for the event rate: p ϵ {0.001, 0.01, 0.1, 1.0}, which represent different levels of control. For example, p = .001 would represent a well-controlled process that rarely experiences instability. We set S_c ϵ {1, 1.5, 2, 2.5, 3, 3.5}. We set N ϵ {1, 2, 3, 4} and k ϵ {2,3}. We varied these factors to produce 576 scenarios (6 levels for S_c, 4 levels for p, and 4 levels for N, 3 levels for g(S|E), and 2 levels for k). We calculated the performance characteristics (sensitivity, specificity, PPV, NPV) for each scenario. We illustrate the method with example calculations for two assays, thyroid stimulating hormone (TSH) and methotrexate (MTX).

Overall, our model is designed to show how a QC plan (a selection of k and N) performs in a particular context (g(S|E), p, S_c). Our model extends the work of Westgard [5]. Westgard determined the predictive values (NPV, PPV) after assuming values for the probability of error detection and false rejection (P_{ed}, P_{fr}). Westgard's work was a useful insight that connected diagnostic test metrics (PPV, NPV) to QC; however, by assuming values for P_{ed} and P_{fr}, Westgard's model fails to make a connection between assay parameters (k, N, g(S|E), S_c, p) and QC performance. Our model provides a direct connection between assay parameters (context) and QC performance. (Fig. 5) The context is defined by the critical shift size, the frequency of upsets, p, and the distribution of shifts, g(S|E). A QC plan is defined by the QC limits, k, and the number of repeats, N. We develop metrics to evaluate QC plan performance in different contexts.

4. Results

We will focus on results for the uniform distribution and a QC limit of k = 2 (i.e., 2σ). As described below, results for other shift distributions (exponential, triangular) and QC policies (i.e., k = 3) were qualitatively similar. Results for all 576 scenarios are provided in an Excel file which is provided as Supplementary Table 1 in the Supplementary materials.

The predictive values varied widely across the different scenarios (Fig. 6). For example, the PPV varied from < 1% to nearly 100%. The PPV decreased as the critical shift size, S_c, increased. The PPV increased

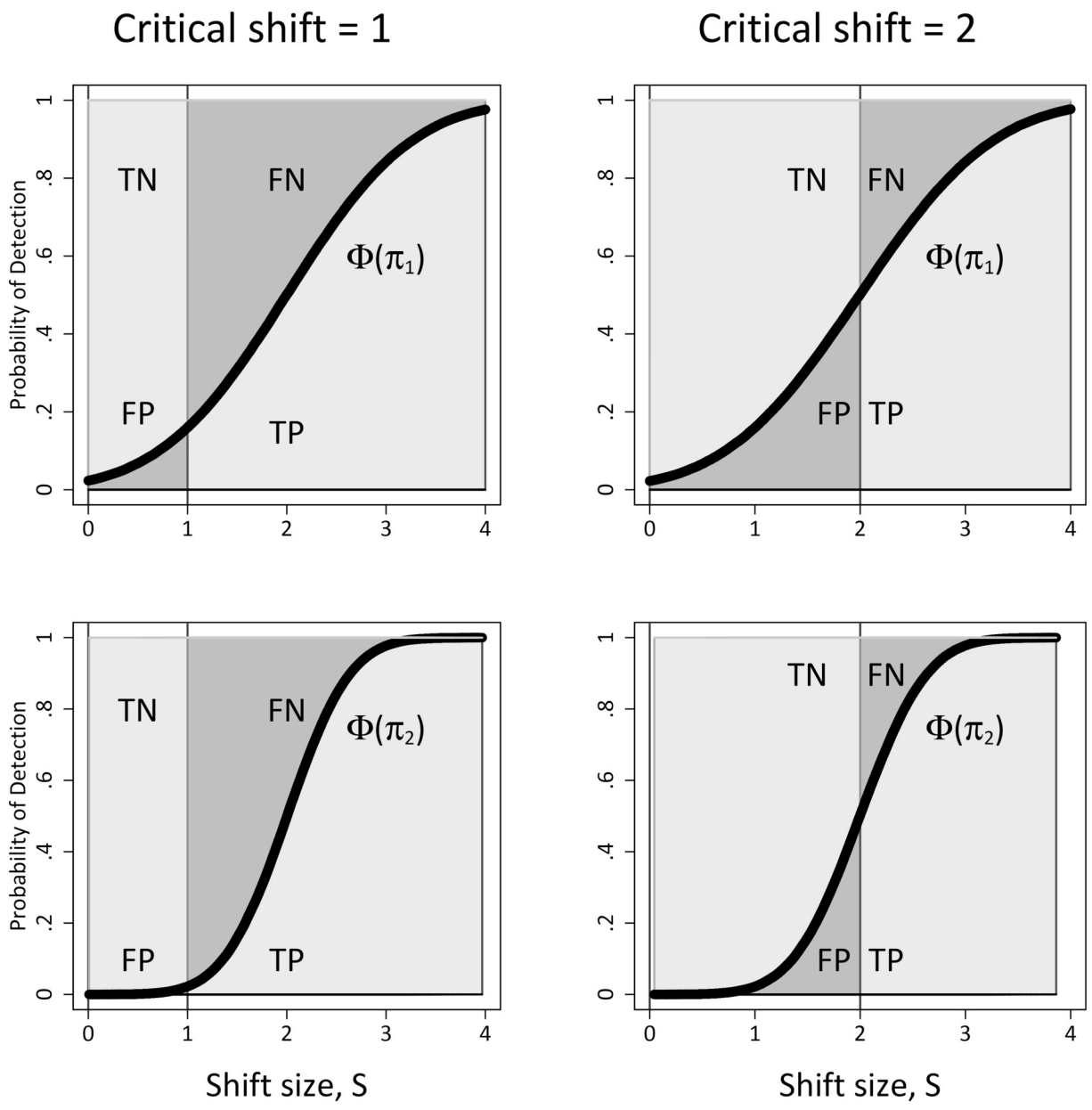

Fig. 3. Impact of Critical Shift Size and Statistical Power on Accuracy of Quality Control (QC). The sloping curves are power curves, Φ, for two different QC policies, $\pi_1$and π_2. The QC policies may differ by the number of repeats taken or the QC rules applied. TN = true negative, FN = false negative, TP = true positive, FP = false positive. Shifts above the critical shift size are considered important, and shifts below the critical shift size are considered unimportant. The QC system is designed to detect shifts in the output distribution. A true positive occurs when there is an important shift and the QC rule fails. A false positive occurs when there is an unimportant shift and the QC rule fails. The length of a line from Φ to 1 and zero to Φ indicate the probabilities of the associated events. The shift size is expressed as multiples of the SD.

with the number of repeats, N. The PPV was most sensitive to N when the critical shift size was low and the probability of a shift event, p, was low. The PPV increased with the probability of a shift event. PPV increased as the ratio k/S_c increased (Supplementary Table 1).

The NPV increased as the critical shift size, S_c, increased (Fig. 6). The NPV increased with the number of repeats, N, and was most sensitive to N when the critical shift size was low and the probability of a shift event, p, was low. The NPV decreased with the probability of a shift event. In general, the NPV was less sensitive than the PPV to changes in the inputs. NPV decreased as the ratio k/S_c increased.

The relationships between the predictive values (PPV and NPV) and the input parameters (N, S_c) were not strongly dependent on the distribution of shifts, g(S|E). The results for the exponential and triangular distribution were qualitatively similar to the results for the uniform distribution (see Supplementary Figs. 1 and 2).

In general, the sensitivity and specificity were less sensitive than the

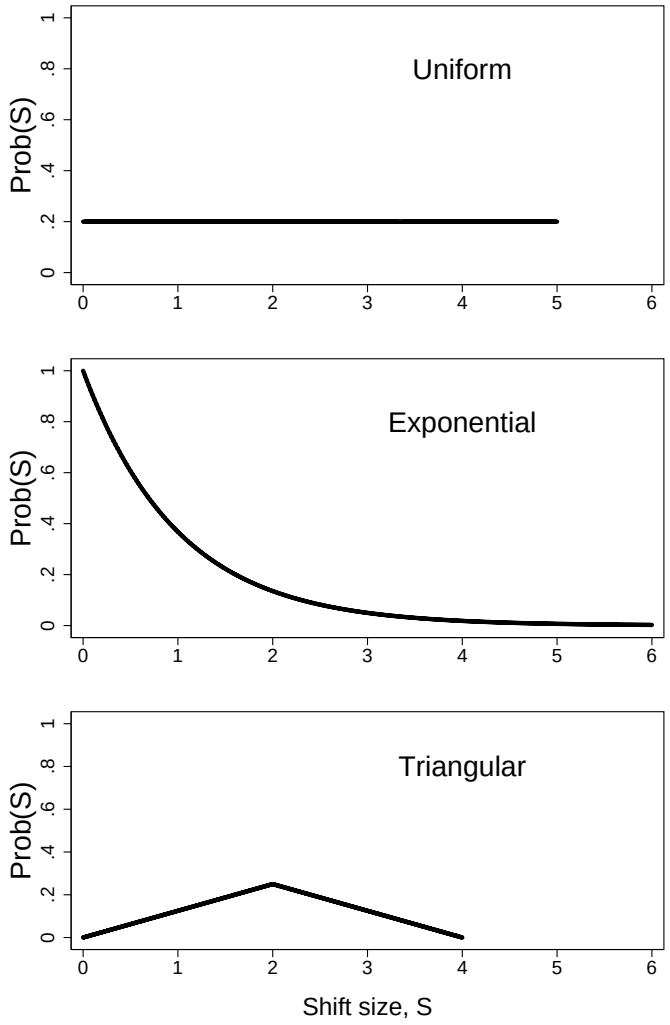

Fig. 4. Probability Distributions Used in the Model. Each distribution shows the probability of a shift size of magnitude S (expressed as multiples of the SD). The uniform distribution models a context in which all shifts (within a specific range) are equally likely. The exponential distribution models a context in which smaller shifts are more likely that large shifts. The triangular distribution models a context in which shifts tend to have a specific magnitude. The horizontal axis is the probability of observing a shift size of magnitude S.

predictive values to changes in inputs (Fig. 7). Sensitivity increased and specificity decreased with the critical shift size, Sc. Sensitivity was relatively insensitive to the event rate p.

5. Example calculations

We performed example calculations for 2 assays, thyroid stimulating hormone (TSH) and methotrexate (MTX). We assumed an event rate, p, of 0.01 and that shift sizes, g(S), were uniformly distributed between 0 and 5 SDs. We used Westgard's definition of a critical shift [6]. We also assumed that the lowest acceptable capability level was 3.0. The observed capabilities (sigma) were 6.2 and 4.7 for TSH and MTX, respectively. The corresponding critical shift sizes were 3.2 (TSH) and 1.7 (MTX). Using a 1–3 s control rule, the sensitivity, specificity, PPV, and NPV were 0.83, 0.99, 0.45, and 1.00 for TSH and 0.59, 1.00, 0.59, and 1.00 for MTX (Table 2).

6. Discussion

A measurement process is subject to random events that affect the accuracy of the results. Some of these events are small and unimportant. Others are significant. Ideally, a QC plan would alert operators to significant events and ignore unimportant events. A QC strategy that could reliably distinguish important shifts from trivial shifts would save time and money. We developed a simple model that places process shifts into two such categories: important and unimportant. These categories are created by defining a critical shift size, S_c, which depends on the assay and medical decision points. This scheme enabled us to apply standard metrics such as sensitivity, specificity, and predictive values that provide insights into QC system performance.

Westgard introduced the idea of a critical shift which is used in the construction of an operations specification (OPSpec) chart [6]. Westgard's definition of a critical shift is based on process capability and is the largest shift in the mean that could be tolerated without producing an unacceptable level of results exceeding the total allowable error. Roughly, the Westgard method defines a critical shit as one that will cause 5% of patient specimens to have measurement error that exceeds a clinically significant limit (e.g. TEA) [7]. This definition focuses on compliance. Our definition is consistent with Westgard's but is more flexible. In our scheme, the critical shift size can be varied depending on the purpose of the QC plan and on medical risk. For example, if one were concerned with compliance, one could adopt Westgard's definition; however, the rate of unacceptable results (i.e. those exceeding TEA) should vary depending on the assay. For some assays a 5% rate of results exceeding TEA may be acceptable because errors have low risk (e.g. diagnoses are made on the basis of multiple tests). In some cases, diagnoses are made on the basis of a single test (e.g. genetic tests for a cancer mutation that directs chemotherapy) and the criteria should be much more stringent. Alternatively, if one were focusing on process improvement, one might select a different critical shift size. Also, our method is not confined to assay results and can be applied to any process that is monitored using statistical process control.

In the traditional power curve analysis, the probability of false rejection occurs only when no shift as occurred (i.e. $S = 0$). In our model, we define a region of unimportant shifts defined as shifts which are smaller than the critical shift size (i.e. $S < S_c$). Our model includes the traditional power curve analysis as a special case by setting S_c to a very small number.

Prior research on QC strategies reasons from cause to effect. For example, given an event (e.g., a shift), one can calculate the likelihood of detecting the event (statistical power). Through a simple application of Bayes Theorem, our model provides a way to reason from effect to cause. For example, given a rule violation, our model can calculate the probability that an important event occurred or that an unimportant event occurred. This type of information supports the types of decisions that operators actually face. An operator does not ask, “Given a shift, what is the probability that I will detect it”? Rather, they respond to signals generated by the QC monitoring system and ask, “Given a signal, what is the probability that it is correct?” “Given a QC signal, is it worth my time to troubleshoot?” Or, “Given no signal, am I confident that the assay is providing reliable results?” Our model provides insight into QC performance from the end-user perspective.

Our results show that the effectiveness (PPV, NPV) of a QC plan depends on the frequency of events and on the types of upsets that are most likely to occur. Any prior information about the pattern of out-of-control behavior could be used to improve the effectiveness of QC. Unfortunately, little is known about out-of-control behavior (i.e., event frequency and the distribution of shift sizes). However, even rough approximations might be sufficient to improve the effectiveness of QC

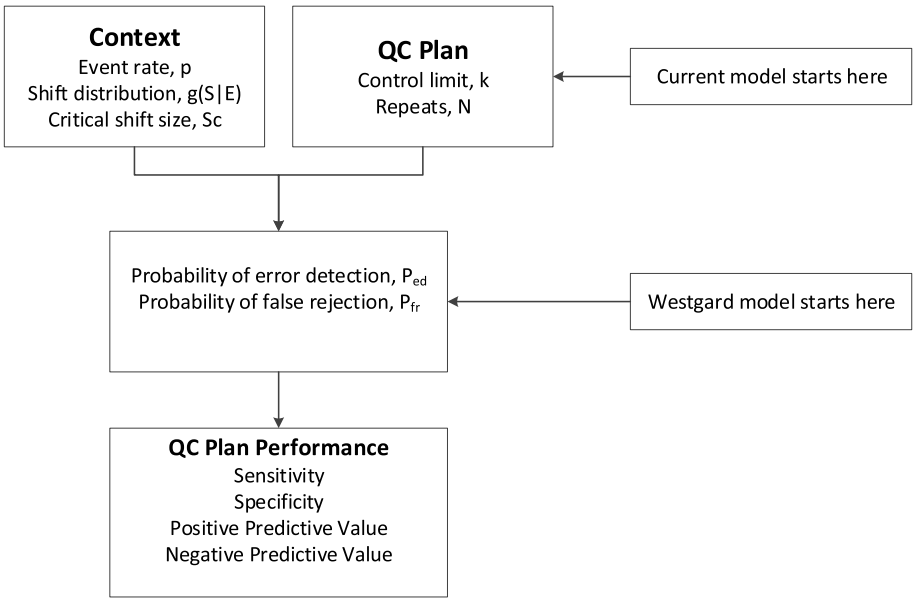

Fig. 5. Comparison of Predictive Models. The Westgard model starts by assuming the probability of error detection and false rejection (P_{ed}, P_{fr}). The model in this paper relates these statistics (P_{ed}, P_{fr}) to attributes of the assay (context) and to decision variables (k, N) that are controlled by the laboratory.

and might be a fruitful area of further research.

Our example calculations show that three sigma limits had reasonable accuracy and predictive values. The specificity and NPV were close to 100% for both assays. The positive predictive values were close to 50%. This is because of the low prevalence (p = .01) of shift events. In practice, this means that rule violations would be generated by important shifts only about 50% of the time. The PPV could be improved by adjusting the control limits upward (e.g., k = 4).

We studied the impact of using multiple repeats (N > 1) at a QC level. In other industries, QC is generally monitored with X bar charts which plot the average of multiple samples at each time point. In clinical chemistry, the most common practice is to monitor an assay with a single QC sample at each level (N = 1). Taking multiple samples increases statistical power but increases cost. We are not aware of any formal economic analysis of QC practice in clinical chemistry but we suspect that the general practice of N = 1 is based on economic considerations. We examined scenarios with N > 1 to explore the potential costs and benefits of increasing N. Increasing N reduces the width of control limits because the SD decreases with the square root of N.

We also explored a wide range of critical shift sizes. While some of these may be impractical (e.g., Sc > 2) the choice of Sc depends on the decision context (compliance vs process improvement) and would be driven by the relative costs of false negatives and false positives. A high value of Sc (e.g. Sc > 2) might be practical for a very capable process. For example, a shift of 2 sigma might not be important in a six sigma process. A recent study reported many assays with capability that exceeded six sigma [8].

Our model examines the ability of a QC plan to detect a discrete change in state (a shift in the mean) at a specific point in time. However, this is not the only type of change that occur. For example, drifts are a common source of error. Our model detects whether the mean has exceeded the critical shift level at any particular point in time. In time, our method would detect a shift due to a trend. It would ignore the small shifts at the beginning of the trend because they are unimportant. Using our model for detection of trends is a potential area of future research.

To our knowledge, the impact of out-of-control behavior has not been considered in prior work. Our results show that the predictive values are highly dependent on the event rate, p. This makes intuitive sense. If the event rate is low (for example, 0.001), then a rule violation is most likely a false positive. On the other hand, if the event rate is high, a rule violation is much more likely to be a true positive. A well-controlled process would have a low event rate and, in those circumstances, the QC policy (choice of N, and k) should be adjusted to reduce false alarms.

Our model suggests that QC plans should be designed according to the upset patterns that are expected to occur. Given the lack of knowledge regarding upset behavior, it is reasonable to ask whether our model can provide any practical insights. In the absence of any knowledge, one might assume that a uniform distribution of shift sizes provides a reasonable approximation of the state of knowledge. One might be able to obtain a rough estimate of the event rate from rule violations. These could be used to calculate PPVs and NPVs. Given these inputs, our model could be used to explore the impact of various QC policies on performance. Thus, even in the absence of specific knowledge of upset behavior, our model can provide some broad guidance (albeit approximate) for adjusting QC limits if one is able to provide a critical shift size. The critical shift size need not be exact. It was necessary to make some assumptions about unknown quantities that, for sake of argument, we regarded to be reasonable.

Our model is limited by the need to make assumptions regarding upset behavior; however, traditional QC plan design requires similar assumptions. As in our model, traditional QC design involves a trade-off between the cost of false positive and false negative results to adjust the QC limits. Researchers have incorporated costs into QC design but these methods are complex and are rarely used in practice. Thus, QC is often based on implicit assumptions regarding the relative costs of false negatives and false positives, as well as the frequency and magnitudes of

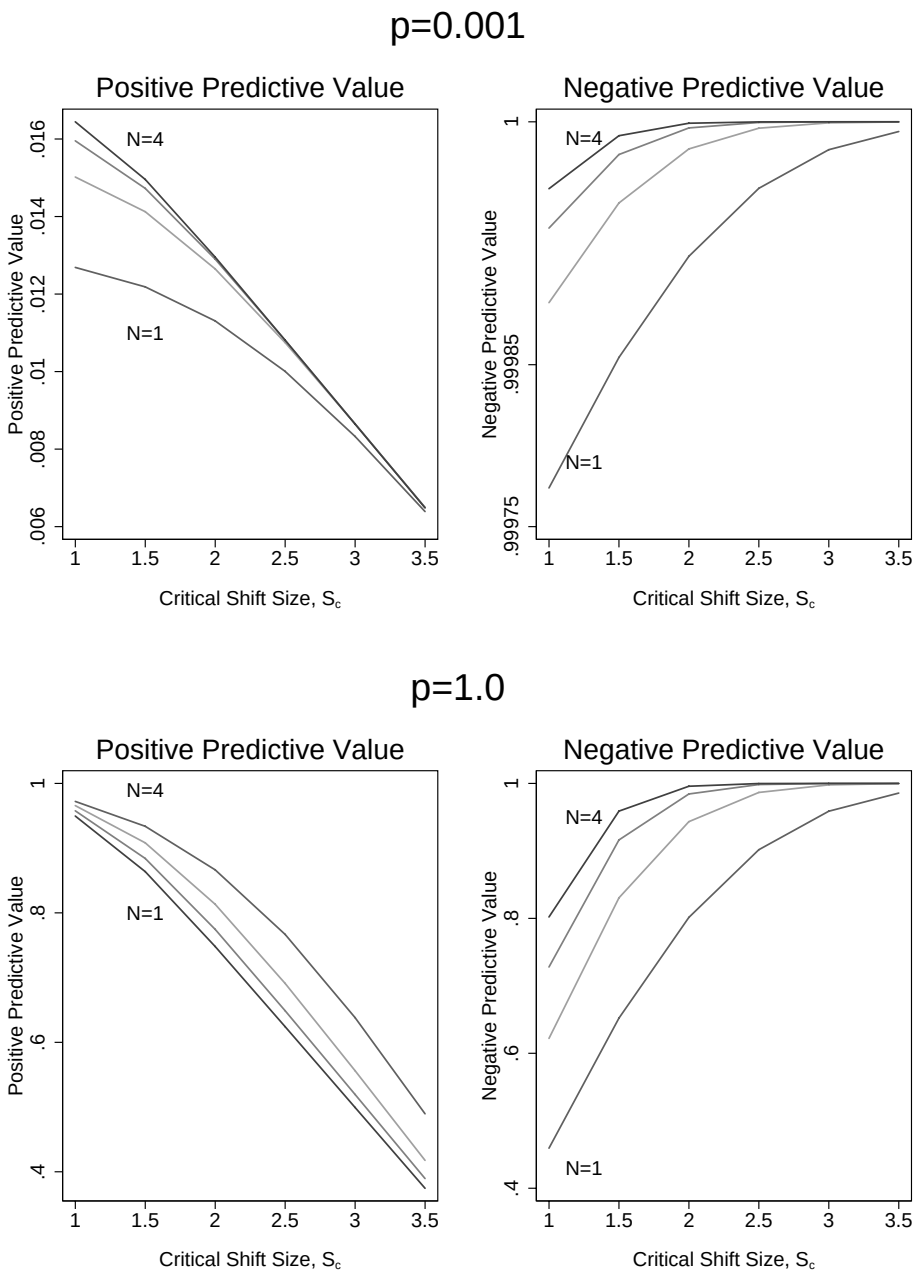

Fig. 6. Predictive Values (uniform distribution, control limit = 2 s). N = the number of repeat samples. p = the probability of a shift event.

those events. For example, a 1–3 s rule implies that a false-positive result bears higher costs than a false-negative result. Our model uses a simple dichotomous cost scheme (important, unimportant) which is relatively easy to apply. In addition, traditional QC design ignores the frequency and magnitude of events. Clearly, these are important factors. Our model makes these assumptions explicit and enables analysts to explore the impact of these assumptions on QC performance.

Our model extends the work of Westgard and Groth [5]. Their model began by assuming values for the probability of error detection and false rejection. Our model derives these quantities based on properties of a QC plan and the context in which it operates. Thus, our model provides a direct connection between basic properties of an assay and QC plan performance.

Future research could be directed at finding methods to estimate upset behavior. Our model uses two parameters to describe upsets: (1) the frequency of upsets, p and (2) the magnitude of the event. This is a very flexible model but there may be better approaches. Also, work on setting the critical shift size could be useful. We have observed that, in practice, operators implicitly employ such rules. For example, a 4–1 s or 10× violation is often considered a warning rather than a criterion for rejection. The 4–1 s and 10× rules identify relatively small shifts, and the fact that these signals are often ignored implies that shifts of this magnitude are relatively unimportant. Our model also suggests the possibility of optimizing QC limits based on the relative costs of false

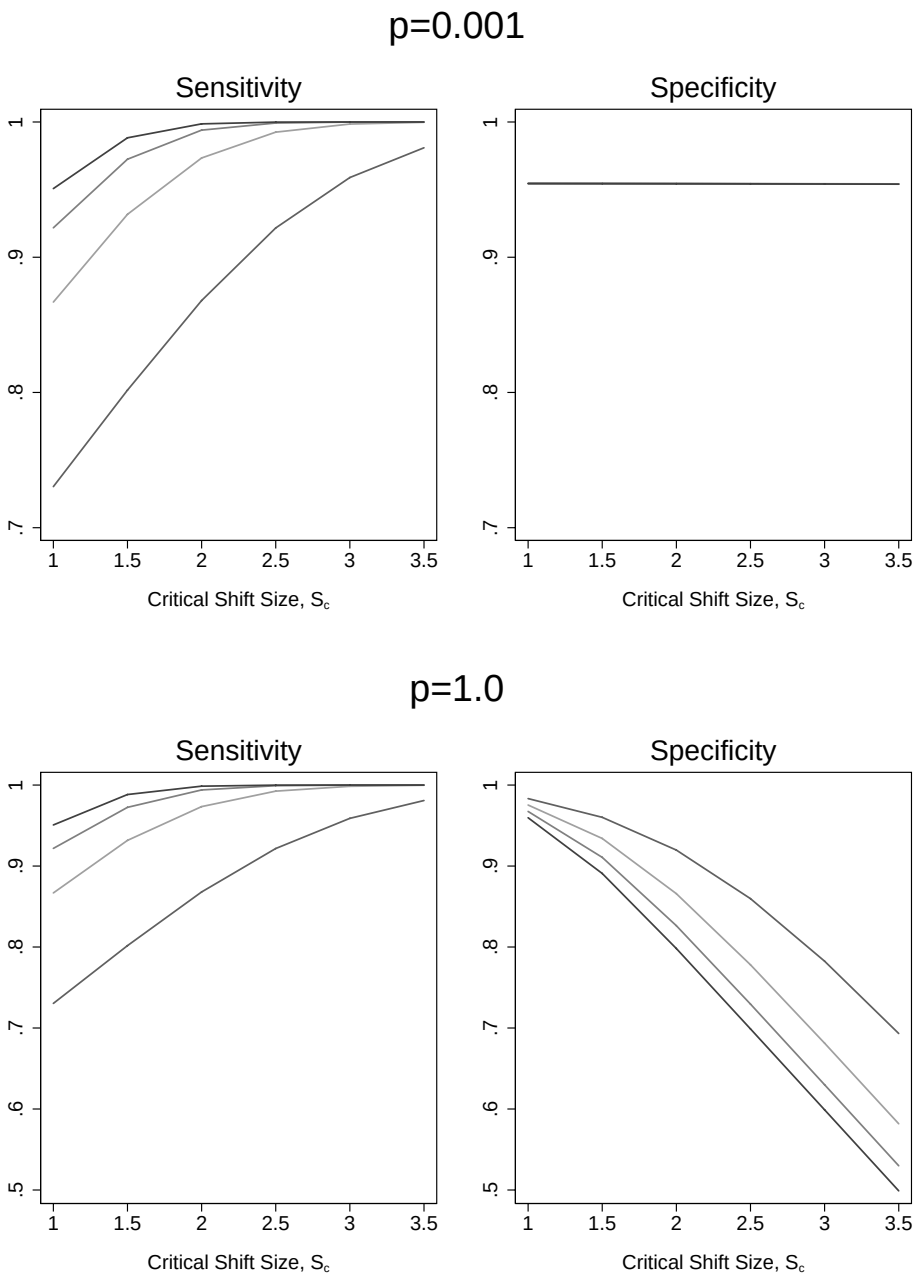

Fig. 7. Sensitivity and Specificity (uniform distribution, control limit = 2 s). Each line represents the number of repeats (N = 1 to 4, N = 4 is the upper line, N = 1 is the lower line). p = the probability of a shift event. The lines in the upper right panel appear to coincide. The sensitivity varies in the third decimal place. This level of variation can't be detected on this scale. Supplementary Fig. 3 shows this panel on a smaller scale.

positives (cost of troubleshooting when a trivial error has occurred) and false negatives (failing to detect an important change in the assay). We pursue this in a related study. Future research might determine whether our approach could be extended to examine other error patterns (e.g. trends) and to study optimal intervals for QC. One could extend our model to determine performance characteristics such as the expected number of QC events between false rejections, expected length of time between false rejections, and the expected number of QC events to detect an out-of-control state. Finally, research on statistical methods to estimate the prevalence and the distribution of the magnitude of errors might be fruitful areas for future research.

In conclusion, we have shown that context affects QC performance and that QC performance could be improved if context were considered in the design of QC policies. We presented a simple model that considers context and allows one to apply common accuracy metrics to evaluate QC performance. In particular, we present metrics (PPV, NPV) that evaluate performance from an end-user perspective. These metrics could enable analysts to improve the effectiveness of QC plans.

Table 2
Example calculations.

Assay	Thyroid stimulating hormone (mIU/L)	Methotrexate (μmol/L)
Mean	25.8	0.88
SD	1.1	0.038
Number of repeats (N)	1	1
Control limit, k	3	3
Event frequency, p	.01	.01
Distribution of shifts	Uniform (0.5)	Uniform (0.5)
Bias	3.3%	5.0%
Total allowable error, TAE	30%	25%
Capability (sigma)	6.2	4.7
Critical shift, S_c	3.2	1.7
Sensitivity	0.83	0.59
Specificity	0.99	1.00
Positive predictive value	0.45	0.59
Negative predictive value	1.00	1.00

Appendix A. Set theory notation

- A∩B is the intersection of two sets and contains the elements that are common to A and B.
- A∪B is the union of two sets and contains the elements in A or B.
- A^c is the complement of A. A^c contains all the elements that are not in A.
- Set membership. For discrete sets, the members are simply listed. For example, A = {1,4,6}. Continuous sets are described using set builder notation which shows the condition which must be satisfied for set membership. For example, A = {S: 0 < S < 10} says that A is composed of all numbers S such that S is between 0 and 10. Continuous sets can also be expressed as interval. For example, x ϵ [0,10) says that x is a number between 0 and 10.

Appendix B. List of symbols

Symbol	Symbol type	Description
A	Event	An important shift has occurred (i.e., $S > S_c$)
A^c	Event	Same as U.
B	Event	An unimportant positive shift has occurred (i.e., $0 < S < S_c$)
E	Event	A positive shift has occurred (i.e., $S > 0$)
E^c	Event	Complement of E. Event where S = 0.
T	Event	QC rule is violated. For example, $\overline{X}_t > k\sigma_{\overline{x}}$
T^c	Event	QC rule is not violated ($\overline{X}_t < k\sigma_{\overline{x}}$).
U	Event	An unimportant shift has occurred (i.e., $0 \leq S < S_c$). U differs from B because U includes shifts of zero.
g(S|E)	Probability distribution	This is the probability distribution of S given that event E has occurred (S > 0).
P(A)	Probability	A probability is associated with an event. This is the probability that event A occurs.
P(U|T)	Conditional probability	This is the probability of event U occurring given that event T has occurred.
k	Input parameter	Size of the QC limit, $k\sigma_{\overline{x}}$.
μ	Variable	The mean of the QC result distribution, X_{it}. $\mu = \mu_0 + S$.
μ_0	Variable	The baseline mean of the QC result distribution (i.e., when S = 0).
n	Input	The number of QC measurements (repeats) taken at each QC event
S	Random variable	Shift size
s_c	Input parameter	The critical shift level. This separates important shifts ($S > S_c$) from unimportant shifts ($S < S_c$).
σ	Input parameter	The SD of the QC result distribution, X_{it}.
X_{it}	Random variable	The i^{th} QC observation at time t.
$\overline{X}_t$	Random variable	The average of the n QC measurements taken at time t.

Appendix C. Derivation of accuracy statistics

We now develop a mathematical model that will allow us to calculate accuracy measures. Let A designate the event that an important shift ($S \geq S_c > 0$) occurred, and let B designate the event that an unimportant shift ($0 < S < S_c$) occurred:

$$A = \{S: S \geq S_c\} \cap E \quad (1)$$

$$B = \{S: S < S_c\} \cap E \quad (2)$$

Note that B only includes positive shifts and does not include shifts of size 0. Clearly, shifts of size 0 are unimportant. Therefore, we define a set of unimportant events, U, which includes both the set of shifts of size 0 and inconsequential positive shifts:

$$U = B \cup E^c = A^c \quad (3)$$

Given these definitions, we can define performance measures as follows:

A true positive is an event in which a significant shift has occurred and the QC system provides an out-of-control signal:

$$TP = (T \bigcap A) \tag{4}$$

A true negative is an event in which a nonsignificant event occurs and the QC system does not provide an out-of-control signal:

$$TN = (T^c \bigcap U) \tag{5}$$

A false positive is an event in which a nonsignificant event occurs and the QC system generates an out-of-control signal:

$$FP = (T \bigcap U) \tag{6}$$

A false negative is an event in which a significant event occurs and the QC system does not generate an out-of-control signal.

$$FN = (T^c \bigcap A) \tag{7}$$

In general, these events are conditional on the sampling policy, π = (k, N), and the SD of the QC values, σ_x.

Probability model: We wish to calculate performance statistics as a function of a QC plan, π. To form a probability model, we calculate probabilities for each of the events defined in the previous section. The probabilities are interpreted as the frequency of occurrence over a large number of QC events. We also calculate performance measures such as the sensitivity, P(T|A); specificity, P(T^c|U); positive predictive value, P(A|T); and negative predictive value, P(U|T^c). The probability of an important shift given that a positive shift (S > 0) occurred is given by:

$$P(A \mid E) = P(S > S_c \mid E) = \int_{S_c}^{\infty} g(S \mid E)dS \tag{8}$$

Given P(A|E), other related probabilities are easily calculated:

$$P(B \mid E) = 1 - P(A \mid E) \tag{9}$$

$$P(B) = P(B \mid E)P(E) \tag{10}$$

$$P(A) = P(A \mid E)P(E) \tag{11}$$

Given P(A), the probability of an unimportant shift (i.e., $0 \le S < S_c$) can be calculated.

$$P(U) = P(A^c) = P(B) + P(E^c) = P(B) + (1 - P(E)) \tag{12}$$

The events, A, B, and E^c are disjoint and partition event T (Fig. 2). Therefore, the probability of a QC rule violation is given by:

$$P(T) = P(T \cap A) + P(T \cap B) + P(T \cap E^c) \tag{13}$$

These components are the probability of the joint occurrence of a QC rule violation (T) and an important positive shift, $P(T \cap A)$, the probability of the joint occurrence of a QC rule violation (T) and an unimportant positive shirt, $P(T \cap B)$, and the probability of the joint occurrence of a QC rule violation (T) and no shift. These probabilities are calculated as follows:

$$P(T \cap A) = P(\mathrm{T} \mid A)P(A) = \int_{s_c}^{\infty} P(\overline{X}_t > k\sigma_{\bar{x}} \mid \mu = \mu_0 + S)g(S \mid E)P(E)dS \tag{14}$$

$$P(T \cap B) = P(T \mid B)P(B) = \int_{0}^{s_c} P(\overline{X}_t > k\sigma_{\bar{x}} \mid \mu = \mu_0 + S)g(S \mid E)P(E)dS \tag{15}$$

$$P(T \cap E^c) = (1 - P(E))P(\overline{X}_t > k\sigma_{\bar{x}} \mid \mu = \mu_0) \tag{16}$$

$$P(T \cap U) = P(T \cap B) + P(T \cap E^c) \tag{17}$$

The function, $P(\overline{X}_t > k\sigma_{\bar{x}}\mu = \mu_0 + S)$, is the power curve for the QC plan. We assume that QC events have a Bernoulli distribution with parameter p.

The sensitivity (probability of error detection, P_{ed}) can be computed from eqs. (4) and (7):

$$\mathrm{Sn} = \mathrm{P_{ed}} = \mathrm{P(T \mid A) = P(T \bigcap A)/P(A)} \tag{18}$$

The specificity is given by

$$\mathrm{Sp = P(T^c \mid U) = 1 - P(T \bigcap U)/P(U)} \tag{19}$$

because P(T^c|U) + P(T|U) = 1. The probability of false rejection, P_{fr}, is 1 – Sp.

The events A (important shift) and U (unimportant shift) partition the sample space. Therefore,

$$\mathrm{P(A \mid T) = 1 - P(U \mid T)} \tag{20}$$

P(A|T) is the probability of a significant shift given a QC failure. This is the PPV of the QC policy. Similarly, applying Bayes Theorem:

$$PPV = P(A \mid T^c) = P(T^c \mid A)P(A)/P(T) \tag{21}$$

Events A and U partition the sample space. Therefore,

$$NPV = P(U \mid T^c) = 1 - P(A \mid T^c) \tag{22}$$

$P(U|T^c)$ is the probability that an unimportant shift occurred and the process is in control. This is the NPV of the QC policy. The quantities Sn, Sp, PPV, and NPV are the key performance parameters of the QC system.

The performance of a QC plan, π, depends on the sampling plan and on the size of the critical shift, S_c (Fig. 3). In our model, the QC plan is defined by the QC limits, $k\sigma_{\bar{x}}$, and the number of repeat samples, N, taken at each QC event. The QC strategy and the imprecision of the assay determine the shape of the power curve.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2019.04.053.

References

[1] CLSI, Statistical Quality Control for Quantitative Measurement Procedures: Principles and Definitions, Clinical Laboratory Standards Institute, Wayne, PA, 2016.
[2] D.C. Montgomery, Introduction to Statistical Quality Control, John Wiley & Sons, New York, 2009.
[3] C.A. Parvin, L. Kuchipudi, J.C. Yundt-Pacheco, Should i repeat my 1:2s QC rejection? Clin. Chem. 58 (5) (2012) 925–929.
[4] J. Westgard, T. Groth, Power functions for statistical control rules, Clin. Chem. 25 (6) (1979) 863–869.
[5] J.O. Westgard, T. Groth, A predictive value model for quality control: effects of the prevalence of errors on the performance of control procedures, Am. J. Clin. Pathol. 80 (1) (1983) 49–56.
[6] J.O. Westgard, P.L. Barry, Basic QC Practices: Training in Statistical Quality Control for Medical Laboratories, Westgard QC, (2010).
[7] D.D. Koch, J.J. Oryall, E.F. Quam, D.H. Feldbruegge, D. Dowd, P.L. Barry, J.O. Westgard, Selection of medically useful quality-control procedures for individual tests done in a multitest analytical system, Clin. Chem. 36 (2) (1990) 230–233.
[8] S. Westgard, V. Petrides, S. Schneider, M. Berman, J. Herzogenrath, A. Orzechowski, Assessing precision, bias and sigma-metrics of 53 measurands of the Alinity ci system, Clin. Biochem. 50 (18) (2017) 1216–1221.