An Analysis of Multirules for Monitoring Assay Quality Control

*Brandon S. Walker, MS, Lauren N. Pearson, DO, MHS, Robert L. Schmidt, MD, PhD, MBA**

Laboratory Medicine 2020;51:94-98

DOI: 10.1093/labmed/lmz038

ABSTRACT

Background: Multirules are often employed to monitor quality control (QC). The performance of multirules is usually determined by simulation and is difficult to predict. Previous studies have not provided computer code that would enable one to experiment with multirules. It would be helpful for analysts to have computer code to analyze rule performance.

Objective: To provide code to calculate power curves and to investigate certain properties of multirule QC.

Methods: We developed computer code in the R language to simulate multirule performance. Using simulation, we studied the incremental performance of each rule and determined the average run length and time to signal.

Results: We provide R code for simulating multirule performance. We also provide a Microsoft Excel spreadsheet with a tabulation of results that can be used to create power curves. We found that the R_{4S} and 10x rules add very little power to a multirule set designed to detect shifts in the mean.

Conclusion: QC analysts should consider using a limited-rule set.

Keywords: quality control, error detection, false rejection, Westgard rules, multirules, simulation

Quality control (QC) is used to monitor processes to detect departures from normal operations or instability. Any change in the distribution of the measured output suggests the presence of assignable cause variation that is synonymous with instability. QC processes use statistical methods to detect a signal (*assignable cause variation*) in the presence of background noise (*common cause variation*).[1]

Various methods are available to monitor QC processes. The simplest and most common rule is to use 3 SD (sigma) limits, classify the process as unstable, and begin troubleshooting when results exceed these limits (*1_{3S} rule*). Although such rules have the advantage of simplicity, they are not very sensitive and often fail to detect small shifts in the mean. For this reason, many laboratories use multirules to increase the sensitivity of QC monitoring.

Multirules are created by applying several different criteria for QC failure. For example, 2 consecutive results on 1 side of the 2-sigma limit might be considered a failure and be used in addition to the 3-sigma limit. The process would fail if either rule were triggered.

Although these rules increase sensitivity, they have several disadvantages. First, each rule increases the probability of false rejection; this probability can be substantial when many rules are applied. Second, it is difficult to predict the performance of the monitoring system in response to rule adjustments. Power curves have been published for a standard set of rules (Westgard rules),[2,3] but the list of potential rules is very long, and it is impractical to publish power curves for each case. It would be helpful if analysts had simple tools that could be used to predict the behavior of rule adjustments.

Power curves can be calculated, but these calculations require advanced mathematical methods (Markov analysis) that are not accessible to most laboratorians. Also, power curves can be generated using simulation, which is much more

Abbreviations

QC, quality control; CUSUMs, cumulative sums; EWMAs, exponentially weighted moving averages; ARLs, average run lengths; 1_{3S}, 1 value exceeding 3 SD; 2_{2S}, 2 consecutive results exceeding 2 SD (both in the same direction); 4_{1S}, 4 consecutive results exceeding 1 SD (all in the same direction); 10x, 10 consecutive results on 1 side of the center line; R_{4S}, 2 consecutive results with 1 greater than 2 SD and 1 less than –2 SD

Department of Pathology, University of Utah and ARUP Laboratories, Salt Lake City

*To whom correspondence should be addressed.
Robert.schmidt@hsc.utah.edu

Table 1. Rule Definitions

Rule	Definition
1_{3S}	Run fails if measurement is 3 SD greater or 3 SD less than the mean
2_{2S}	Run fails if 2 consecutive measurements are 2 SD greater than or 2 SD less than the mean
4_{1S}	Run fails if 4 consecutive measurements are 1 SD greater than or 1 SD less than the mean
10x	Run fails if 10 consecutive measurements are greater than or 1 SD less than the mean
R_{4S}	Run fails if a measurement is 4 SD greater than the mean and the next measurement is 4 SD less than the mean, or vice versa

accessible. The published power curves were produced by simulation; however, to our knowledge, the code was not published. Analysts could easily experiment with QC rule adjustments if such code were available. Thus, the objective of this study was to provide code to calculate power curves and to investigate certain properties of multirule QC.

Materials and Methods

The probability of QC failures was estimated through simulation, using the statistical software R with the R package Propagate.[4, 5] We selected R because it is widely used and freely available. The R code is available in the Appendix. Briefly, we simulated the probability of failure of Westgard rules after shifts of 1 to 6 SDs. For each shift, we generated 1,000,000 measurements and then applied the Westgard rules to estimate the probability of failure.

Rule definitions are provided in **Table 1**. For multirules, which involve more than 1 rule, failure was triggered if any rule failed. For example, multirule $1_{3S}/2_{2S}$ could be triggered if measurement fails 1_{3S} or 2_{2S} rules.

We calculated the power curve for each individual rule as a function of shift size. Also, we calculated the probability that each rule would fail when combined in a rule set.

Results

The results of our simulation are presented in **Figure 1**; also, those results are tabulated in Supplementary Table 1.

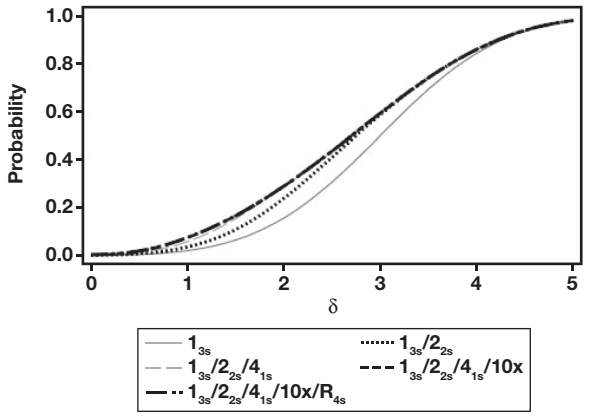

Figure 1

Power curves for combinations of sensitizing rules. The graph is the detection rate for a single quality-control (QC) level. Delta indicates the shift size in SDs; 1_{3S}, 1 value exceeding 3 SD; 2_{2S}, 2 consecutive results exceeding 2 SD (both in the same direction); 4_{1S}, 4 consecutive results exceeding 1 SD (all in the same direction); 10x, 10 consecutive results on 1 side of the center line; R_{4S}, 2 consecutive results with 1 greater than 2 SD and 1 less than –2 SD.

Supplementary Table 1 provides the probability of detecting a shift for a wide range of shift sizes and sampling plans (ie, repeat levels).

The error-detection rate for each individual rule is presented in **Table 2**. Individual rules are generally insensitive to small shifts (ie, shifts smaller than 2 SD). Combined rules increase the error-detection rate for small shifts (**Table 3**) but also increase the probability of false rejection. For example, a 1_{3S} rule has a 15.91% probability of detecting a 2 SD shift in the mean and has a false-rejection rate of 0.28% (**Table 2**). The full set of Westgard rules (final column, **Table 3**) has a 29.13% chance of detecting a 2 SD shift, but the false-rejection rate (ie, the probability of a rule failure when the shift size is 0) is 0.63%. The R_{4S} rule does not increase the statistical power for detecting shifts but does increase the probability of false rejection (**Tables 2** and **3**).

When rules are combined and applied in the order of 1_{3S}, 2_{2S}, 4_{1S}, 10x, and R_{4S}, errors are most often detected by the 1_{3S} and 2_{2S} rules (**Table 4**). The 4_{1S}, 10x, and R_{4S} rules contribute very little to the statistical power for error detection (**Table 3**).

DOI: 10.1093/labmed/lmz038

Table 2. The Probability of Error Detection by Westgard Rules and Size of Shift[a]

Size Shift	Rule				
	1_{3S}	2_{2S}	4_{1S}	10x	R_{4S}
0 SD	.28%	.09%	.11%	.10%	.10%
1 SD	2.28%	2.17%	3.33%	3.41%	.04%
2 SD	15.91%	16.65%	15.93%	8.78%	0%
3 SD	49.95%	38.42%	23.58%	9.92%	0%
4 SD	84.11%	48.30%	24.91%	10.00%	0%
5 SD	97.70%	49.89%	25.00%	10.00%	0%

1_{3S}, 1 value exceeding 3 SD; 2_{2S}, 2 consecutive results exceeding 2 SD (both in the same direction); 4_{1S}, 4 consecutive results exceeding 1 SD (all in the same direction); 10x, 10 consecutive results on 1 side of the center line; R_{4S}, 2 consecutive results with 1 greater than 2 SD and 1 less than −2 SD.
[a]The results are for 1 quality control (QC) level. Each rule is considered separately. For example, a 2_{2S} rule (applied alone with no rules) would have a 16.65% chance of detecting a shift of 2 SDs.

Table 3. Probability that a QC Run with 1 Level Will Fail, by Westgard Rule Combination and Size Of Shift[a]

Size Shift	Rule Set				
	1_{3S}	$1_{3S}/2_{2S}$	$1_{3S}/2_{2S}/4_{1S}$	$1_{3S}/2_{2S}/4_{1S}$/10x	$1_{3S}/2_{2S}/4_{1S}$/10x/ R_{4S}
0 SD	.27%	.36%	.46%	.55%	.63%
1 SD	2.28%	3.82%	6.02%	7.69%	7.71%
2 SD	15.84%	24.04%	28.90%	29.12%	29.13%
3 SD	50.01%	58.62%	59.26%	59.26%	59.26%
4 SD	84.12%	85.73%	85.74%	85.74%	85.74%
5 SD	97.72%	97.77%	97.77%	97.77%	97.77%

QC, quality control; 1_{3S}, 1 value exceeding 3 SD; 2_{2S}, 2 consecutive results exceeding 2 SD (both in the same direction); 4_{1S}, 4 consecutive results exceeding 1 SD (all in the same direction); 10x, 10 consecutive results on 1 side of the center line; R_{4S}, 2 consecutive results, with 1 greater than 2 SD and 1 less than −2 SD.

Table 4. Distribution of QC (1-Level) Failures when All Westgard Rules are Applied[a]

Shift Size	Rule					
	1_{3S}	2_{2S}	4_{1S}	10x	R_{4S}	TOTAL
0 SD	41.76%	13.56%	16.33%	12.79%	15.56%	100%
1 SD	30.46%	20.47%	27.30%	21.43%	0.35%	100%
2 SD	54.52%	28.14%	16.52%	0.83%	0.01%	100%
3 SD	84.26%	14.67%	1.07%	0.00%	0.00%	100%
4 SD	98.01%	1.99%	0.00%	0.00%	0.00%	100%
5 SD	100.00%	0.00%	0.00%	0.00%	0.00%	100%

QC, quality control; 1_{3S}, 1 value exceeding 3 SD; 2_{2S}, 2 consecutive results exceeding 2 SD (both in the same direction); 4_{1S}, 4 consecutive results exceeding 1 SD (all in the same direction); 10x, 10 consecutive results on 1 side of the center line; R_{4S}, 2 consecutive results, with 1 greater than 2 SD and 1 less than −2 SD.
[a]Given a shift of 3 SD, the 1_{3s} rule is triggered 84% of the time, the 2_{2S} rule is triggered 15% of the time, and the 4_{1S} rule is triggered 1% of the time.

The run-length distribution (ie, the number of QC events before a rule failure) is skewed to the right, and the average run length decreases with the size of the shift (**Figure 2**). For example, the number of events between false rejections (shift size = 0) ranges from 0 to 1500, whereas the number of events required to detect a shift of 3 SD ranges from 0 to 4.

Discussion

We calculated power curves for Westgard multirules. Our results match those of other researchers.[3] Also, we provided tabulated results for a wide range of shift sizes and repeat levels. The tabulated results provide a convenient way to

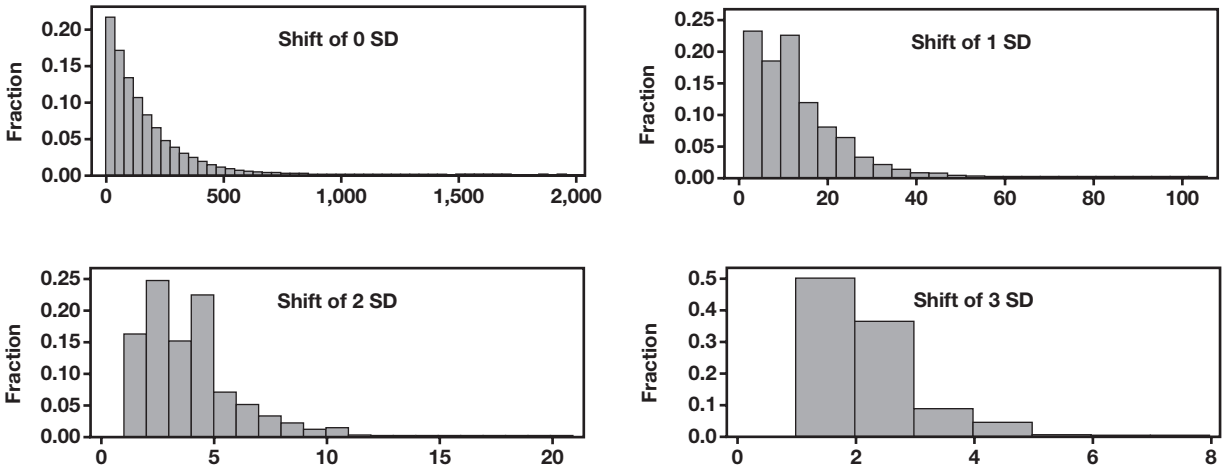

Figure 2

The run-length distribution for all Westgard rules. The *run length* is the number of quality control (QC) events before a rule failure occurs. The graph shows results for a single QC level. A rule failure when the shift size is 0 (upper left panel) is a *false rejection*. The graphs show that the run-length distribution is skewed to the right and the average run length decreases with the shift size. Note that the horizontal scale is different on each graph.

look up probabilities of error detection, a way that is easier than interpolating from figures.

We provided R code for our simulations. The R code enables analysts to verify our calculations and to modify the code, to conduct experiments to see how rule adjustments would affect QC performance. R is freely available and is widely used in the statistical community—we chose to implement our simulations in R so that our code would be widely accessible.

We made several observations about the performance of multirule QC. Additional rules increase the ability of the QC monitoring process to detect small shifts, but extra rules also increase the probability of false rejection. The increase in sensitivity is modest. For example, the combination of the 1_{3S} and 2_{2S} rules provides an 8% increase in the probability of detecting a 2 SD shift, relative to the 1_{3S} rule only (16% vs 24%). As shown in **Table 3**, the application of additional rules has diminishing returns. For example, using all 5 rules increases the probability of detecting a 2 SD shift to 29% but also increases the probability of false rejection to 0.63%. Thus, there is a tradeoff between the improvements in shift detection and false rejection. Some statisticians recommend against using multirules because other monitoring methods, such as cumulative sums (CUSUMs) or exponentially weighted moving averages (EWMAs), can provide better error detection rates with a lower false-rejection rate.[1, 6] However, to our knowledge, the CUSUMs and EWMAs methods are rarely used by clinical laboratories.

Calculations of the false-rejection rate are based on a very narrow definition of false rejection. False rejection occurs only if a QC rule is violated when the shift size is 0. Many small shifts are inconsequential; troubleshooting small shifts is unlikely to be productive. Thus, the formal definition of false rejection underestimates the rate of unproductive rule violations. It might be more reasonable to categorize shifts as important and unimportant using a critical-shift threshold. For example, if the critical shift size was 1 SD, a rule set with all 5 rules would have a 7.71% chance of false rejection (**Table 3**). This rate of false rejection is unacceptably high.

We focused on detection of shifts. Our results show that the R_{4S} rule should not be included in a rule set designed to detect shifts. The R_{4S} rule may be effective for detecting changes in dispersion. However, when included in a rule set designed to detect shifts, the R_{4S} rule only increases the probability of false rejection without increasing the probability of shift detection (**Tables 2** and **3**).

QC practice in clinical laboratories differs from practice in other industries. In other industries, control charts are used to monitor dispersion (R or s charts) and location (X charts or Xbar charts).[1,7,8] Clinical laboratories do not use charts to

DOI: 10.1093/labmed/lmz038

monitor dispersion.[9] The R_{4S} rule might be a useful component in a multirule approach for monitoring changes in dispersion; however, we are not aware of multirule approaches for monitoring dispersion. Thus, it is unclear whether R_{4S} rule has any usefulness in the multirule sets used in clinical laboratories.

Similarly, we question whether the 10x rule is useful. According to our calculations, it contributes very little to the statistical power for error detection yet adds to the rate of false rejection. In our experience, many laboratories use the 10x rule as a warning and do not initiate troubleshooting when the rule is violated. This practice is problematic. In a regulated environment, there should be clear rules regarding signals and responses.

We did not explore the performance when rules are applied across multiple QC levels. For example, a 2_{2S} rule could be classified as violated when results from 2 separate QC levels are higher than 2 SD. Clearly, cross-level rules would increase the false-rejection rate but, depending on the correlation between levels, cross-level rules could also increase the error detection rate. Multiple levels should be monitored with multivariate QC. However, although this approach has been adopted by other industries, it has not been adopted by clinical laboratories—this is a subject for future research.

Average run lengths (ARLs) are generally used to evaluate QC performance in other industries. The *run length* is the number of trials before a rule failure occurs. We presented data on the run-length distribution for various shift sizes. Our data show that run-length distributions are highly variable, particularly when shift sizes are small. Thus, for small shift sizes, the time to detection can be highly variable. The run length distribution is an important consideration in a QC plan. Our R code can be used to explore the impact of rule selection on the run-length distribution.

Our study provides insight into the performance of multirules and provides computer code that can be used to further explore such rules. Although multirules increase the ability to detect shifts, we question whether this is the optimal approach. Also, it is unclear whether the traditional set of multirules is optimal. Ideally, rules would be optimized based on the relative cost of false rejections and failure to detect errors.

Exploring modifications to multirules is difficult. Although we have provided code that can be easily modified to conduct experiments to determine the impact of modified rules, this modification may be difficult for many analysts. Thus, we question whether using multirules is the optimal approach. In conclusion, our study provides insights into the performance of multirule QC and provides tools that analysts can use to further explore the performance of multirule QC. **LM**

References

1. Montgomery DC. *Introduction to Statistical Quality Control.* New York, NY: John Wiley & Sons; 2009.
2. Westgard J, Groth T. Power functions for statistical control rules. *Clinical Chemistry.* 1979;25(6):863–869.
3. Parvin CA, Kuchipudi L, Yundt-Pacheco JC. Should I repeat my 1:2s QC rejection? *Clinical Chemistry.* 2012;58(5):925–929.
4. R Core Team. The R Project for Statistical Computing. https://www.R-project.org. Accessed on May 16, 2019.
5. Spiess A-N. Package 'propagate'. https://cran.r-project.org/web/packages/propagate/propagate.pdf. Accessed on May 16, 2019.
6. Allison Jones-Farmer L, Woodall WH, Steiner SH, Champ CW. An overview of Phase I analysis for process improvement and monitoring. *J Qual Technol.* 2014;46(3):265–280.
7. Wheeler DJ, Chambers DS. *Understanding Statistical Process Control*. Knoxville, TN: SPC Press; 2010.
8. Ryan TP. *Statistical Methods for Quality Improvement.* Hoboken, NJ: John Wiley & Sons; 2011.
9. Schmidt R, Pearson L. Quality control limits: are we setting them too wide? *Clin Chim Acta.* 2018;486(11):329–334.