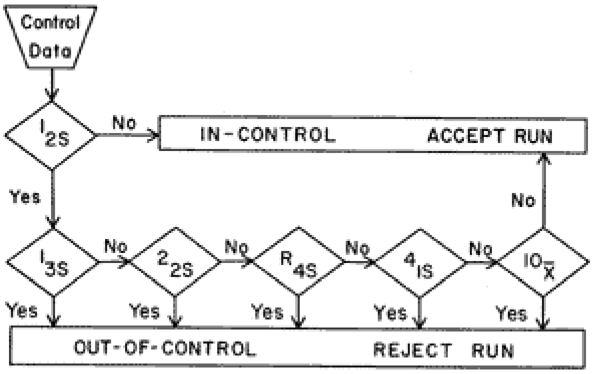
JAMES WESTGARD
FOUNDER

WESTGARD RULES" AND MULTIRULES

What is a multirule QC procedure?

First, a non-technical description. When my daughter Kristin was young and living at home, she liked to party. One day when she told me she was again intending to be out late, I felt the need to exert some parental control over her hours. So I told her that if she was out **once after three, twice after two, or four times after one**, she was in big trouble. That's multirule control.

Kristin hates it when I tell this story, and while it isn't entirely true, it's still a good story and makes multirule QC understandable to everyone. (By the way, she turned out okay; she graduated number 1 in her class from law school and I'm very proud of her. It's also true that she has her mother's brains, which together with my persistence - or stubborness, as it's known around the house - makes a pretty good combination.) I will also have to admit that around our house it is Mrs. Westgard's rules that really count. My wife Joan hates it when I tell this part of the story, but she's put up with me for over thirty years and I'm now in a state of fairly stable control, so it will take a bigger deviation than this before I get into big trouble.



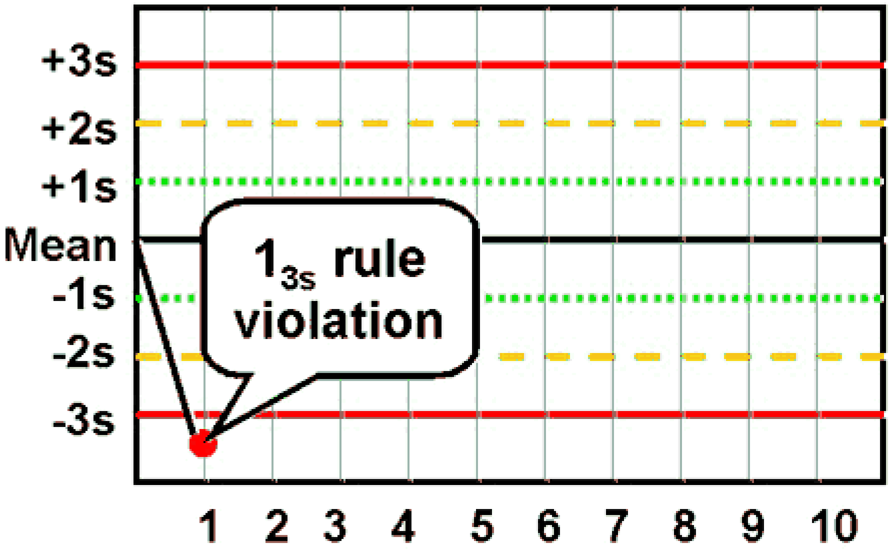
Now for a more technical description. Multirule QC uses a combination of decision criteria, or control rules, to decide whether an analyticalrun is in-control or out-of-control. The well-known Westgard multiruleQC procedure uses 5 different control rules to judge the acceptabilityof an analytical run. By comparison, a single-rule QC procedure uses a single criterion or single set of control limits, suchas a Levey-Jennings chart with control limits set as either themean plus or minus 2 standard deviations (2s) or the mean plusor minus 3s. "Westgard rules" are generally used with 2 or 4 control measurements per run, which means they are appropriate when two different control materials are measured 1 or 2 timesper material, which is the case in many chemistry applications. Some alternative control rules are more suitable when three controlmaterials are analyzed, which is common for applications in hematology, coagulation, and immunoassays.

What are the "Westgard rules"?

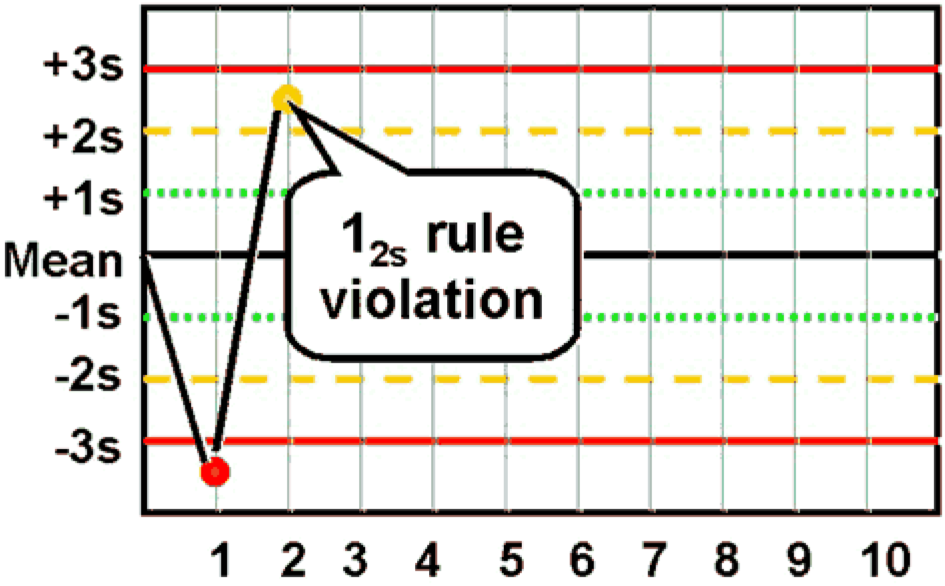
For convenience, we adopt a short hand notation to abbreviate different decision criteria or control rules, e.g., 1_{2s} to indicate 1 control measurement exceeding 2s control limits. We prefer to use subscripts to indicate the control limits, but other texts and papers may use somewhat different notation (e.g. 1:2s rather than 1_{2s}) Combinations of rules are generally indicated by using a "slash" mark (/) between control rules, e.g. $1_{3s}/2_{2s}$.

The individual rule are defined below. The "thumbnail" graphic next to a rule shows an example of control results that violate that rule. You can click on a graphic to get a larger picture that more clearly illustrates the application of each control rule.

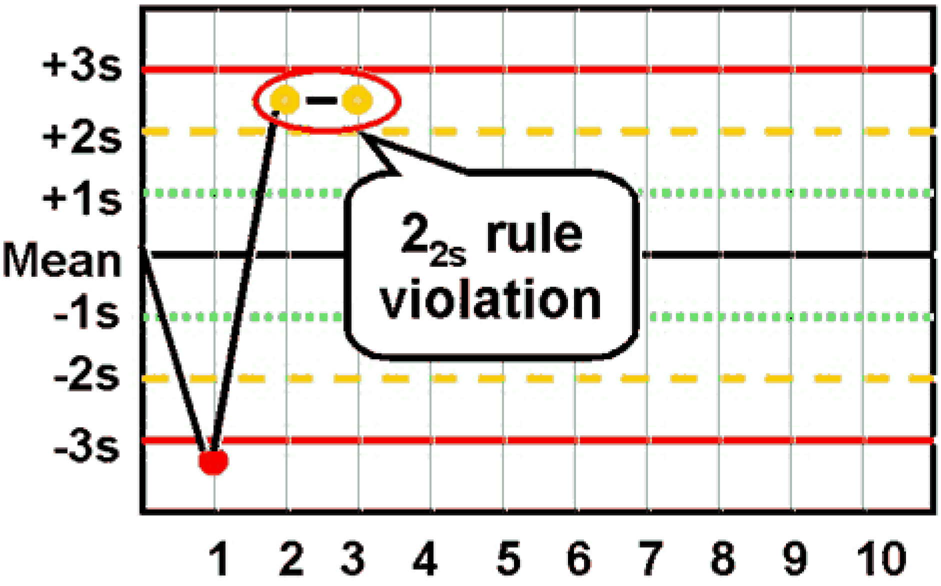
1_{3s} refers to a control rule that is commonly used with a Levey-Jennings chart when the control limits are set as the mean plus 3s and the mean minus 3s. A run is rejected when a single control measurement exceeds the mean plus 3s or the mean minus 3s control limit.



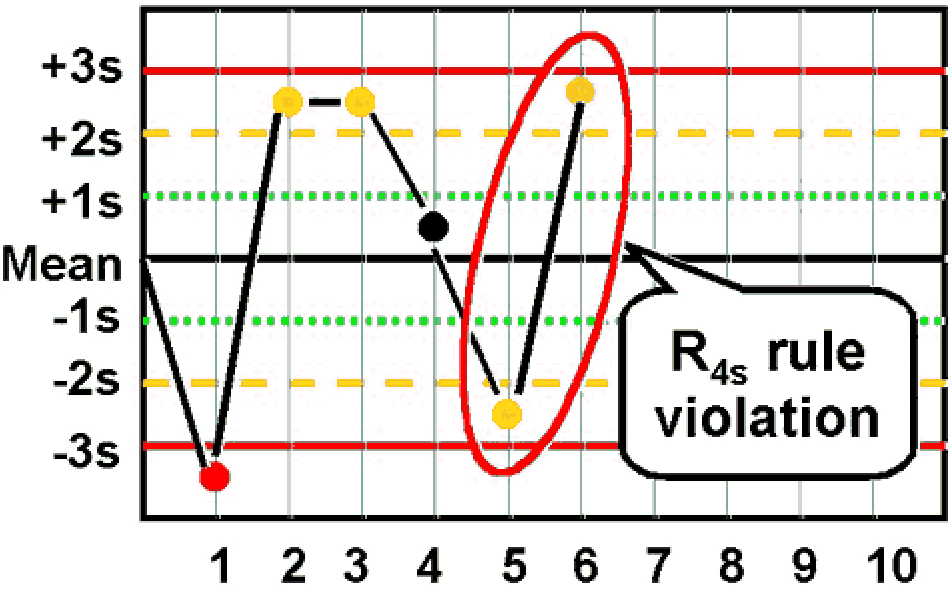
1$_{2s}$ refers to the control rule that is commonly used with a Levey-Jennings chart when the control limits are set as the mean plus/minus 2s. In the original Westgard multirule QC procedure, this rule is used as a warning rule to trigger careful inspection of the control data by the following rejection rules.



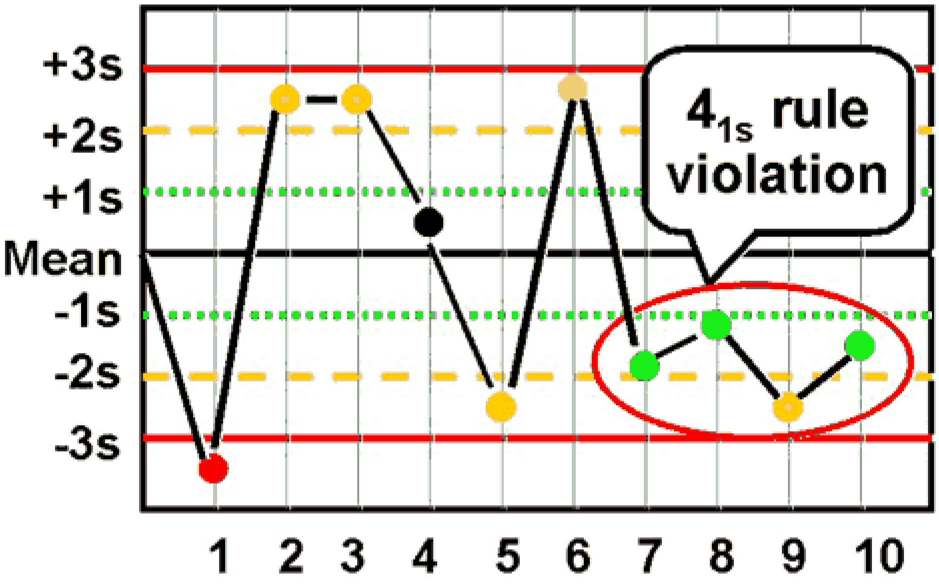
2$_{2s}$ - reject when 2 consecutive control measurements exceed the same mean plus 2s or the same mean minus 2s control limit.



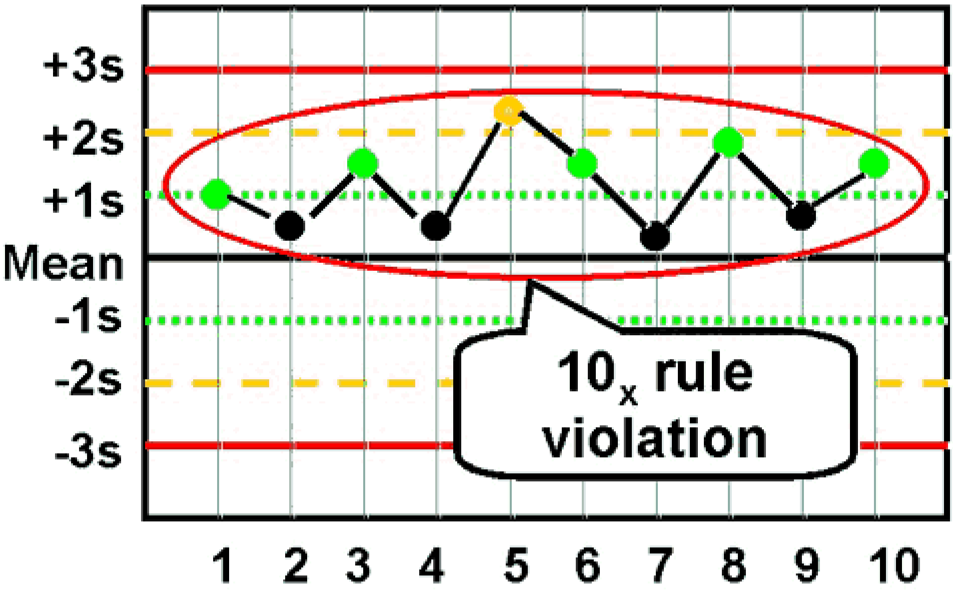
R_{4s} - reject when 1 control measurement in a group exceeds the mean plus 2s and another exceeds the mean minus 2s.



4_{1s} - reject when 4 consecutive control measurements exceed the same mean plus 1s or the same mean minus 1s control limit.

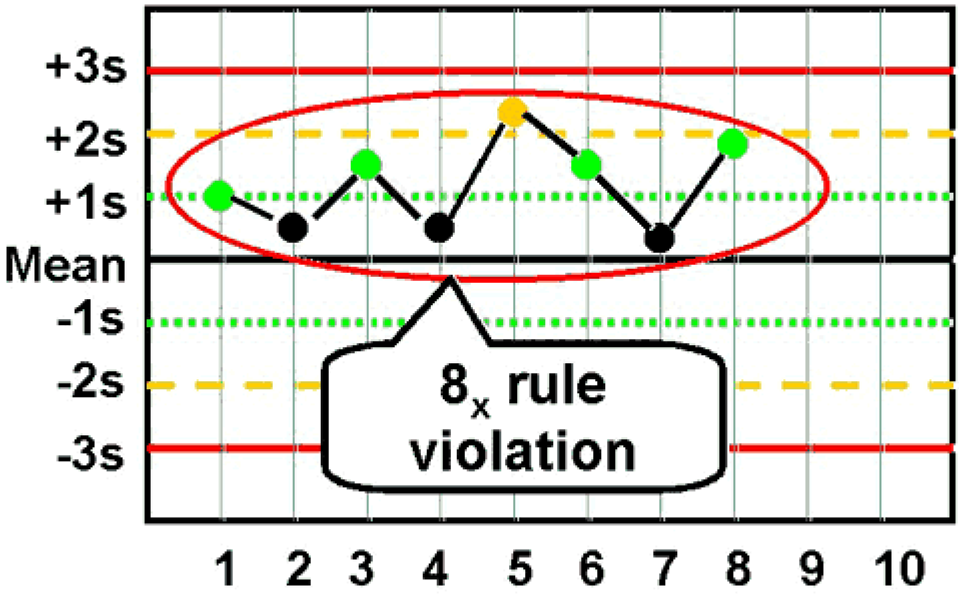


10_x - reject when 10 consecutive control measurements fall on one side of the mean.

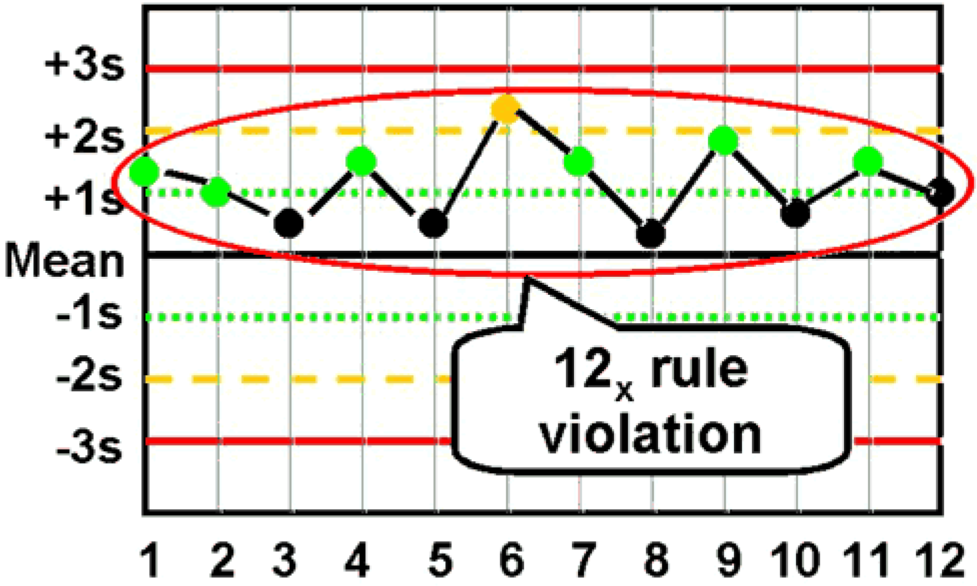


In addition, you will sometimes see some modifications of this last rule to make it fit more easily with Ns of 4:

8_x - reject when 8 consecutive control measurements fall on one side of the mean.



12_x - reject when 12 consecutive control measurements fall on one side of the mean.

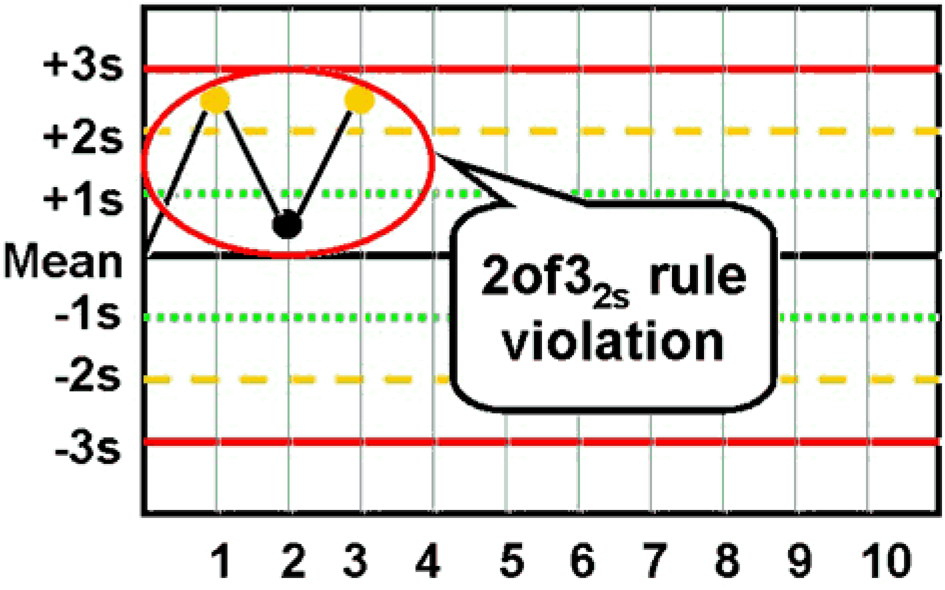


The preceding control rules are usually used with N's of 2or 4, which means they are appropriate when two different controlmaterials are measured 1 or 2 times per material.

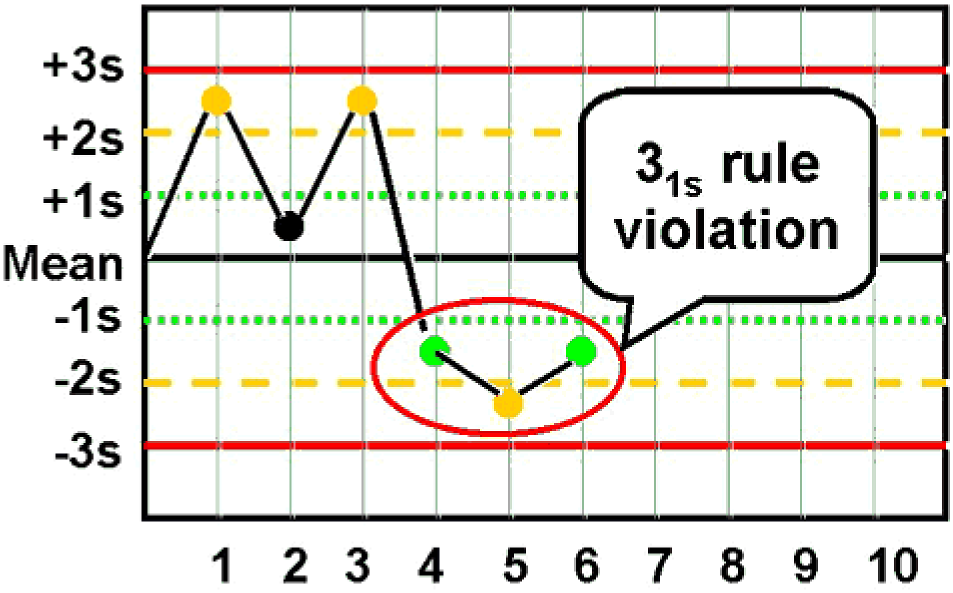
What are other common multirules?

In situations where 3 different control materials are being analyzed, some other control rules fit better and are easier to apply, such as:

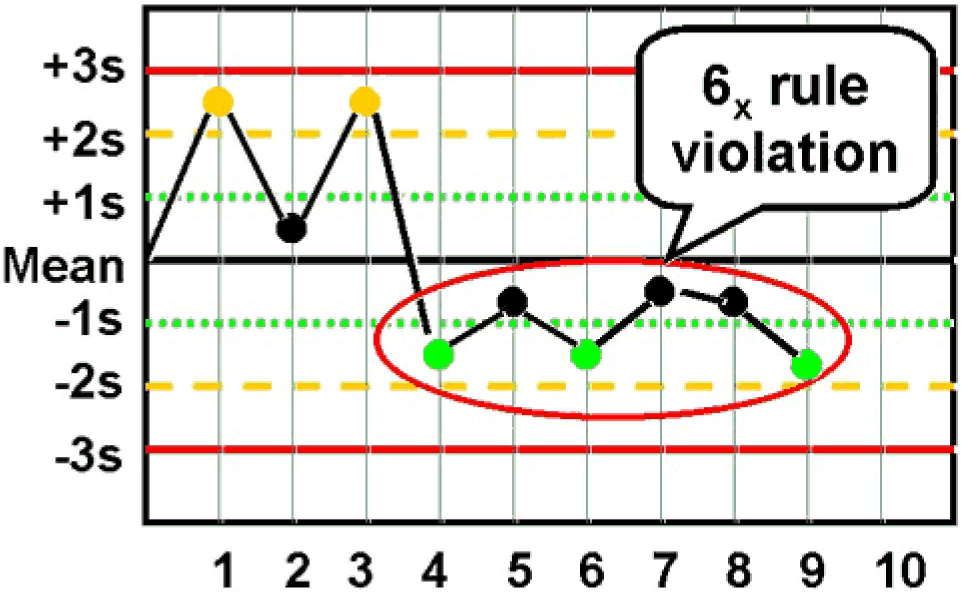
$2of3_{2s}$ - reject when 2 out of 3 control measurements exceed the same mean plus 2s or mean minus 2s control limit;



3_{1s} - reject when 3 consecutive control measurements exceed the same mean plus 1s or mean minus 1s control limit.

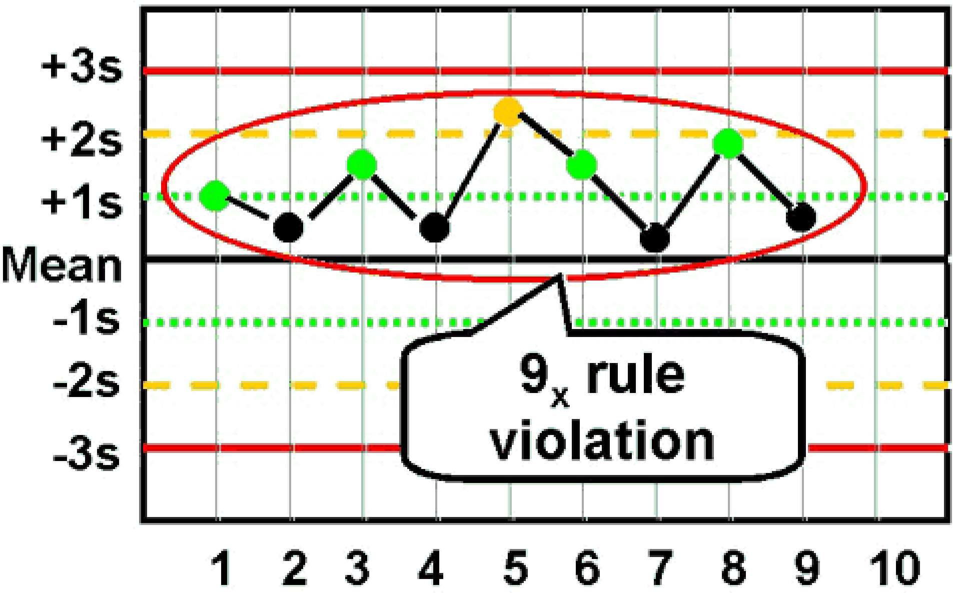


6_x - reject when 6 consecutive control measurements fall on one side of the mean.



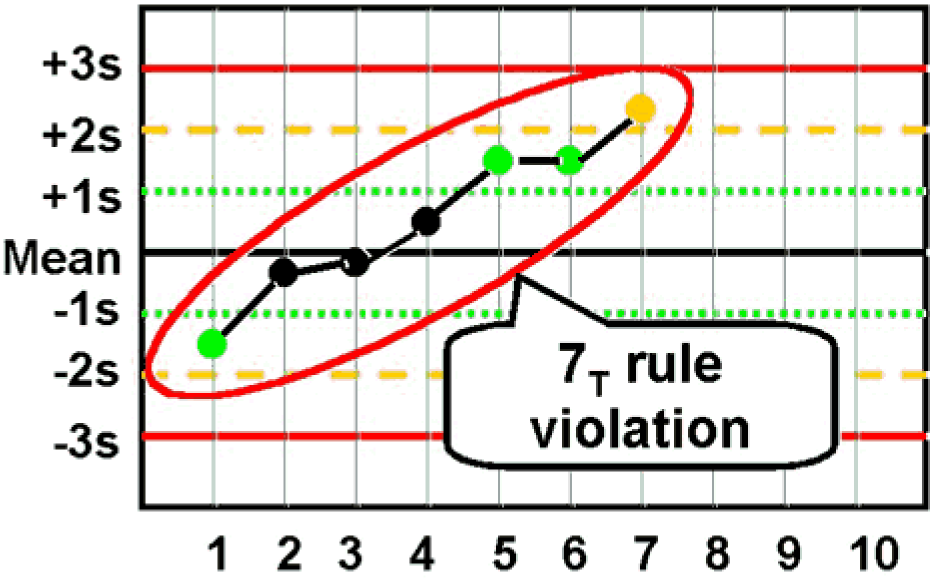
In addition, you will sometimes see some modification of this last rule to include a larger number of control measurements that still fit with an N of 3:

9$_x$ - reject when 9 consecutive control measurements fall on one side of the mean.



A related control rule that is sometimes used, particularly in Europe, looks for a "trend" where several control measurements in a row are increasing or decreasing:

7$_T$ - reject when seven control measurements trend in the same direction, i.e., get progressively higher or progressively lower.



How do you perform multirule QC?

You collect your control measurements in the same way as you would for a regular Levey-Jennings control chart. You establish the means and standard deviations of the control materials inthe same way. All that's changed are the control limits and the interpretation of the data, so multirule QC is really not that hard to do! For manual application, draw lines on the Levey-Jennings chart at the mean plus/minus 3s, plus/minus 2s, and plus/minus 1s. See QC - The Levey Jennings chart for more information about preparing control charts.

In manual applications, a 1_{2s} rule should be usedas a warning to trigger application of the other rules, thus anytimea single measurement exceeds a 2s control limit, you respond by inspecting the control data using the other rules. It's like awarning sign at the intersection of two roads. It doesn't mean stop, it means look carefully before proceeding.

How do you "look carefully"? Use the other control rules to inspect the control points. Stop if a single point exceeds a 3s limit. Stop if two points in a row exceed the same 2s limit. Stop ifone point in the group exceeds a plus 2s limit and another exceedsa minus 2s limit. Because N must be at least 2 to satisfy US CLIA QC requirements, all these rules can be applied within a run. Often the 4_{1s} and 10_x must be used across runs in order to get the number of control measurements needed to apply the rules. A 4_{1s} violation occurs whenever 4 consecutive points exceed the same 1s limit. These 4 may be from one control material or they may also be the last 2 points from a high level control material and the last 2 points from a normal level control material, thus the rule may also be applied across materials. The 10_x rule usually has to be applied across runs and often across materials.

Computer applications don't need to use the 1_{2s}warning rule. You should be able to select the individual rejection rules on a test-by-test basis to optimize the performance of the QC procedure on the basis of the precision and accuracy observed for each analytical method and the quality required by the test.

What is N?

When N is 2, that can mean 2 measurements on one control material or 1 measurement on each of two different control materials. When N is 3, the application would generally involved 1 measurement on each of three different control materials. When N is 4, that could mean 2 measurements on each of two different control materials, or 4 measurements on one material, or 1 measurement on each of four materials.

In general, N represents the total number of control measurements that are available at the time a decision on control status is to be made.

Why use a multirule QC procedure?

Multirule QC procedures are obviously more complicated than single rule procedures, so that's a disadvantage. However, they often provide better performance than the commonly used 1_{2s} and 1_{3s} single-rule QC procedures. There is a false-alarm problem with a 1_{2s} rule, such as the Levey-Jennings chart with 2s control limits; when N=2, it is expected than 9%of good runs will be falsely rejected; with N=3, it is even higher,about 14%; with N=4, it's almost 18%. That means almost 10-20%of good runs will be thrown away, which wastes a lot of time and effort in the laboratory. While a Levey-Jennings chart with 3s control limits has a very low false rejection rate, only 1% or so with Ns of 2-4, the error detection (true alarms) will also be lower, thus the problem with the 1_{3s} control rule is that medically important errors may not be detected. (See QC - The Rejection Characteristics for more information about the probabilities for error detection and false rejection.)

The advantages of multirule QC procedures are that false rejectionscan be kept low while at the same time maintaining high error detection. This is done by selecting individual rules that have very low levels of false rejection, then building up the error detection by using these rules together. It's like running two liver function tests and diagnosing a problem if either one of them is positive. A multirule QC procedure uses two or more statistical tests (control rules) to evaluate the QC data, then rejects a run if any one of these statistical tests is positive.

Are there similiar strategies for QC testing and diagnostic testing?

Yes, a QC test is like a diagnostic test! The QC test attempts to identify problems with the normal operation of an analyticaltesting process, whereas the diagnostic test attempts to identifyproblems with the normal operation of a person. Appropriate actionor treatment depends on correctly identifying the problem.

Both the QC test and the diagnostic test are affected by thenormal variation that is expected when there are no problems,i.e., the QC test attempts to identify changes occurring beyondthose normally expected due to the imprecision of the method,whereas the diagnostic test attempts to identify changes beyondthose normally expected due to the variation of a population (thereference range or reference interval for the test) or the variationof an individual (intra-individual biological variation). Thepresence of this background variation or "noise" limitsthe performance of both the QC test and the diagnostic test.

Are there similar performance characteristics for QC and diagnostic tests?

This background variation causes false alarms that waste timeand effort. These false alarms are more properly called falsepositives for a diagnostic test and false rejections for a QCtest, but both are related to a general characteristic called "test specificity." True alarms are called true positives for a diagnostic test and are referred to as error detection for a QC test, and both are related to a general characteristic called "test sensitivity." Sensitivity and specificity, therefore, are general performance characteristics that can be applied to an test that classifies results as positive or negative (as for a diagnostic test) or accept or reject (for a QC test).

Diagnostic tests are seldom perfectly sensitive and perfectly specific! Therefore, physicians have developed approaches and strategies to improve the performance of diagnostic tests. One approach is to adjust the cutoff limit or decision level for classifying a test result as positive or negative. Both sensitivity and specificity change as this limit changes and improvements in sensitivity usually come with a loss of specificity, and vice versa.

QC procedures, likewise, seldom perform with perfect error detection and no false rejections. Laboratories can employ similar approaches for optimizing QC performance. Changing the control limit is like changing the cutoff limit, and improvements in sensitivity usually come at a cost in specificity (the 1_{2s} rule is an example). Wider control limits, such as 2.5s, 3s, and 3.5s lead to lower error detection and lower false rejections.

How do you use multiple tests to optimize performance?

Another approach for optimizing diagnostic performance is to use multiple tests. To improve sensitivity, two or more tests are used together and a problem is identified if any one of the tests is positive - this is parallel testing. To improve specificity, a positive finding from a sensitive screening test can be followed up with a second more specific test to confirm the problem - this is serial testing. Both sensitivity and specificity can be optimized by a multiple testing approach, but again these changes usually affect both characteristics.

Strategies with multiple tests can also be used to optimize the performance of a QC procedure. Multirule QC is the general approach for doing this. The objectives are to reduce the problems with the false alarms or false rejections that are caused by the use of 2s control limits, while at the same time improving error detection over that available when using

3s control limit. The multiple tests are different statistical tests or different statistical control rules, and the strategies are based on serial and parallel testing.

- False alarms are minimized by using the 1_{2s} rule as a warning rule, then confirming any problems by application of more specific rules that have a low probability of false rejection (serial testing).
- True alarms or error detection are maximized by selecting a combination of the rules most sensitive to detection of random and systematic errors, then rejecting a run if any one of these rules is violated (parallel testing).

When should you use a multirule QC procedure?

Not always! Sometimes a single rule QC procedure gives you all the error detection needed while at the same time maintaining low false rejections. This generally means eliminating the 1_{2s} rule because of its high false rejections and considering others such as $1_{2.5s}$, 1_{3s}, and $1_{3.5s}$ which have acceptably low false rejection rates. The remaining issue is whether adequate error detection can be provided by these other single rule QC procedures. If medically important errorscan be detected 90% of the time (i.e., probability of error detection of 0.90 or greater), then a single rule QC procedure is adequate. If 90% error detection can not be provided by a single rule QC procedure, then a multirule QC procedure should be considered. In general, you will find that single rule QC procedures are adequate for your highly automated and very precise chemistry and hematology analyzers, but you should avoid using 2s control limits or the 1_{2s} control rule to minimize waste and reduce costs. Earlier generation automated systems and manual methods will often benefit from the improved error detection of multirule QC procedures.

To figure out exactly when to use single rule or multirule QC procedures, you will need to define the quality required for each test, look at the precision and accuracy being achieved by your method, then assess the probabilities for false rejection (P_{fr}) and error detection (P_{ed}) of the different candidate QC procedures. Aim for 90% error detection (P_{ed} of 0.90 or greater) and 5% or less false rejections (P_{fr} of 0.05 or less). With very stable analytical systems that seldom have problems, you may be able to settle for lower error detection,say 50%. (See QC - The Planning Process for practical approaches to select appropriate single rule and multirule QC procedures.)

BEST PRACTICES FOR "WESTGARD RULES"

Written by James O. Westgard, Ph.D.

So we've catalogued some of the worst abuses of "Westgard Rules." What about the best uses? What's the best way to use "Westgard Rules" - and When, Why, How and Who, too? Here is a list of 12 practices to make your use of "Westgard Rules" better.

1. Define the quality that is needed for each test.
2. Know the performance of your method (CV, bias).
3. Calculate the Sigma-metric of your testing process.
4. Relate the QC procedure for the test to the Sigma-performance of the method.
5. Use single-rule QC procedures and minimum number of control measurements (N) for methods with high performance
6. Use single-rule QC procedures and moderate number of control measurements (N) for methods with moderate to high performance
7. Use multirule QC procedures for methods with moderate to low performance
8. Use multistage QC designs for methods with undesirable performance
9. Built and interpret multirules in a logical order and adapt the rules to fit with different Ns
10: Define explicitly the application and interpretation of rules within and across matreials and runs.
11. Only use multirules for which error detection characteristics are known.
12. Interpret multirules to help indicate the occurrence of random error or systematic error.
Conclusion

In a previous discussion, we described some of the "abuses, misuses, and in-excuses" involving the improper implementation and interpretation of "Westgard Rules" by instruments, LIS devices, and data workstation QC software. Now that we've cleared the air about the "worst practices", it's time to talk about "best practices" for doing QC right.

It's important to understand the problems (worst practices) in order to implement proper solutions (best practices). If your QC software is doing things wrong, no amount of effort on your part can correct for those problems. QC needs to be done right from the start.

1. Define the quality that is needed for each test.

Quality management begins with the knowledge of the quality that needs to be achieved. Sounds simple, doesn't it? But when I ask the laboratory professionals "What quality is needed for a test?" the answer is seldom a numeric or quantitative definition of the quality requirement. That number could be in the form of a **total allowable error (TE_a)**, such as the CLIA proficiency testing criteria for acceptable performance. Or that number could be in the form of a **clinical decision interval (D_{int})**, which is a gray zone of interpretation for patient treatment. This number comes from the physician and uses his/her diagnosis cutoffs as a way to figure out the level of quality needed in a method. A third possibility for that number is the **biologic total error**, as documented by a European group that has derived figures for the allowable bias and allowable imprecision from studies of individual biological variation. In any case, the sources of some of these numbers are here on the website or somewhere in your laboratory or hospital. Quality begins with defining the quality needed for each test.

If you don't know the quality that is needed, then it doesn't make any difference how you do QC. It's all arbitrary! It's like taking a trip without knowing the destination. Or playing soccer without marking a goal. Or trying to call someone without knowing their phone number - you may get to talk to someone, but they may not care to talk to you.

Resources:

The need for quality standards
CLIA proficiency testing criteria
Clinical decision intervals & quality requirements
European biologic goals
Desirable Precision, Bias and Total Error derived from Biologic Variation (Ricos database)

2. Know the performance of your method (CV, bias).

It's hard to argue with this, too, particularly since CLIA **requires** that a laboratory validate the performance of its methods. You estimate method precision (CV) and accuracy (bias) by method validation experiments when you introduce any new method. For existing methods, the results observed on control materials being analyzed in your laboratory right now can be used to estimate the method's CV and results from proficiency testing or peer comparison studies can be used to estimate bias.

Why is this important? You need to know how well your method is performing. CV and bias are the characterstics that tell you how your method is performing.

Resources:

The comparison of methods experiment - to estimate inaccuracy
The replication experiment - to estimate imprecision

3. Calculate the Sigma-metric for your testing process.

It's useful to have a metric that tells you out-front whether or not your method performance is good enough to achieve the quality that is required. Why do you need to know this? If method performance is bad, no amount of QC can overcome the inherent lack of quality. If method performance is extremely good, only a little QC is needed to assure the necessary quality will be achieved.

Here's the calculation:

Sigma = (TE_a - bias)/CV

Where TE_a is the CLIA allowable total error (expressed in %),
Bias is the systematic error (also expressed in %) compared to a reference method or compared to peer methods in a proficiency testing survey or peer comparison survey, and
CV is the imprecision of your method (in %) as calculated from control measurements in your laboratory.

Here's an example. The CLIA criterion for acceptable performance for cholesterol is 10%. If a laboratory method shows a bias of 2.0% on proficiency testing surveys and a CV of 2.0% on internal QC results, the Sigma-metric is 4 [(10-2)/2]. What does that 4 mean? read on...

Resources:

Six Sigma quality management and desirable precision
Six Sigma calculator

4. Relate the QC procedure for the test to the sigma performance of the method.

The sigma metric will give you a good idea of the amount of QC that is needed. If you have low bias and a small CV, the metric will be high (e.g., TE_a of 10%, bias of 1.0%, and CV of 1.5% gives a sigma of 6.0). Instinctively you know that good method performance should require less QC. If you have a high bias and a large CV (e.g., bias of 3.0 and CV of 3.0 gives a sigma of 2.33), poor method performance will require more QC.

One direct consequence of this practice is that it moves you away from blanket application of any rules for all the tests in the laboratory. You should no longer use just one control rule or one set of control rules on all the tests in your laboratory. You adjust the rules and number of control measurements to fit the performance of the method. The imperative for manufacturers is to provide QC software the flexibility that allows users to optimize QC design on a test by test basis.

Resources:

Six Sigma quality management and requisite QC

For our regular readers, these first four points shouldn't come as a surprise. Since Westgard Web came online in 1996, the articles, lessons, applications, guest essays - pretty much everything we post - have been trying to drive home the point that we need to define the quality we need and measure the performance of our methods. The OPSpecs chart and the Normalized OPSpecs charts (available online for free) are graphic tools to illustrate what control rules are best for your tests. The Validator® and EZ Rules® 3 software programs are automated tools to help you pick the control rules ("Westgard" or otherwise) needed by your tests. Indeed, these first four points are really universal guidelines for establishing the "best practices" for QC. Whether or not you're using the "Westgard Rules" in your laboratory, you need to do these things.

5. Use single-rule QC procedures and minimum Ns for methods with high performance.

Amazingly enough, if method performance is good in relation to the quality needed for the test, you may not need to use multirule QC at all. When sigma is 6.0 or greater, any QC will do; use a simple single-rule procedure with 3.5s or 3.0s control limits and the minimum number of control measurements (typically Ns of 2 or 3). When sigma is 5.5 to 6.0, use 3.0s control limits and Ns of 2 or 3.

6. Use single-rule QC procedures and moderate Ns for methods with moderate to high performance.

For methods having sigmas between 4.5 and 5.5, you need to be more careful in your selection of QC procedures. At the high side (5.0 to 5.5), you can generally use 2.5s control limits with an N of 2 or 3. At the low side (4.5 to5.0), you should use an N of 4.

7. Use multirule QC procedures for methods with moderate to low performance.

This is the corollary to best practices 5 and 6. When single-rule procedures can't provide the high error detection that is needed, then you switch to multirule procedures with Ns from 4 to 6. For method sigmas around 4.0, multirule QC is the way to go.

8. Use multidesign QC procedures for methods with minimum performance.

For method performance in the 3.0 to 3.5 sigma range, you need to do a maximum amount of QC to achieve the necessary error detection. That amount of QC will be expensive, so to minimize the costs, you can adopt two different QC designs - one a STARTUP design for high error detection and the other a MONITOR design for low false rejections. You use the STARTUP during your (you guessed it) startup or any time when the instrument has gone through a significant change. For example, after trouble-shooting and fixing problem, use your STARTUP design to make sure everything is ok again. The idea is to switch back and forth between these designs as appropriate. The STARTUP design should be a multirule QC procedure with the maximum N that is affordable (N=6 to 8). The MONITOR design can be a single rule procedure with a minimum N of 2 or 3.

Multidesign QC is the latest advance in "Westgard Rules". Most QC software programs don't have this capability because manufacturers (a) don't realize that some of laboratory tests perform so badly they need extra QC or (b) don't have the technical expertise in QC to know what QC features to offer their customers. Customers are also to blame because (a) they're happy to do the minimum QC to be in compliance with government regulations, (b) they're often too busy to worry about doing QC correctly, and (c) they're not asking manufacturers for better QC technology. This last reason is why marketing and sales departments in the major diagnostics manufacturers routinely downgrade the priority of QC features in new products. They're listening to the "wants" of their customer, but not addressing the true needs of the customer.

Resources:

Formulating a Total Quality Control Strategy
What is N?
Sage Advice on "New" Approaches to QC

9. Build and interpret multirules in a logical order and adapt the rules to fit with different Ns.

Contrary to public opinion, "Westgard Rules" doesn't mean a single combination of rules, such as the well-known

$$\mathbf{1_{2s}/2_{2s}/R_{4s}/4_{1s}/10_{x}}$$

multirule procedure. That's just the first example we published of a multirule QC procedure. Other combinations are possible. There's no single "Westgard Rule" - which is one of the reasons why we've always preferred the term "multirule QC" over "Westgard Rules."

For certain types of tests, notably hematology, coag, and blood gas, controls tend to be run in three's, i.e., one low control, one middle control, and one high control. For situations like this, it isn't practical to use the "classic Westgard Rules"; those rules were built for controls in multiples of 2. So when you're running 2, 4, 8 controls, use the "classic" rules. When you're running 3 or 6 controls, use a set that works for multiples of threes:

$$\mathbf{1_{3s}/2of3_{2s}/R_{4s}/\ 3_{1s}/12_{x}}$$

10. Define explicitly the application and interpretation of rules within and across materials and runs.

Do you know what it means to apply a control rule ***within-material***, ***across-material***, ***within-run***, and ***across-run***? All of these applications of the multirule give you another chance to detect errors.

If you're running two controls per run, each control on a different level, measuring those controls once (N=2) and using the classic rules, here are the following questions that can come up. If you use the 2_{2s} rule, how do you do it? Are you applying it across materials, so if the first control is out 2s and the second control is also out 2s, you interpret that as a violation? Are you applying it within-material across-runs, so that if in the previous run, the high control was out 2s, and again in this run, the high control was out 2s, is that a violation of the rule?

It gets even more complicated with the larger rules. If you're using two controls and measuring once, how do you interpret the 10_x rule? Do you look-back 10 runs on each control? Do you look back 5 runs on both controls? What if you're running 3 controls, measuring once (N=3) and are working with the 12_x rule? Do you look-back on the last 6 results of two controls, the last 4 of all three controls, or just each control by itself, the last 12 runs?

Most trouble-some of all is the R_{4s} rule. This is a range rule that is meant only to be applied within-run, so it can pick up random errors. If you apply it across-runs, the rule will also detect systematic errors and confuse the identification of random errors.

There are valid reasons to interpret the control rules one way or another. We're not even suggesting there is a complete "right" way to do the interpretation. If you want the 4_{1s} rule to only detect within-material (but across-run), that's fine. Just make sure you spell that out, both in your design, your implementation, and when you explain the features to the customer. If you don't specify what type of interpretation you're going to do, the customer may assume you're doing something more or less than you're doing.

In a way, this is where "manual" interpretation of the "Westgard Rules" is easier than computer implementation. Visually, you can look at the charts and instantly take in the within-run, across-run, within-material, across-material details, and you can choose to disregard any or all of them if you want. A computer program must be explicitly told how to interpret the rules. It won't look-back on previous runs, look across -materials within-run, or within-material across-runs, unless you write the code to do so.

Resources:

The Multirule Interpretation
What's a Run?

11. Only use multirules for which the error detection characteristics are known.

The "Westgard Rules" aren't completely mix and match. You can't use all possible combinations of control rules. You can immediately see that using a $2_{2s}/2\text{of}3_{2s}$ makes no sense at all. What about a 2_{2s} all by itself? Is that even useful?

Here is a table of all the multirule combinations whose rejection characteristics are known:

For Ns of 2 and 4	

$1_{3s}/2_{2s}$	N=1, R=1;
	N=4, R=1

$1_{3s}/2_{2s}/R_{4s}$	N=2, R=1;
	N=4, R=1

$1_{3s}/2_{2s}/R_{4s}/4_{1s}$	N=2, R=2;
	N=4, R=1

$1_{3s}/2_{2s}/R_{4s}/4_{1s}/8_x$	N=2, R=4;
	N=4, R=2

$1_{3s}/2_{2s}/R_{4s}/4_{1s}/10_x$	N=2, R=5;
	N=4, R=3

$1_{3s}/2_{2s}/R_{4s}/4_{1s}/12_x$	N=4, R=3

For Ns of 3 and 6	

$1_{3s}/2\text{of}3_{2s}$	N=3, R=1;
	N=6, R=1

$1_{3s}/2\text{of}3_{2s}/R_{4s}$	N=3, R=1;
	N=6, R=1

$1_{3s}/2\text{of}3_{2s}/R_{4s}/\ 3_{1s}$	N=3, R=1;
	N=6, R=1

$1_{3s}/2\text{of}3_{2s}/R_{4s}/\ 3_{1s}/6_x$	N=3, R=2;
	N=6, R=1

$1_{3s}/2\text{of}3_{2s}/R_{4s}/\ 3_{1s}/9_x$	N=3, R=3;
	N=6, R=2

$1_{3s}/2\text{of}3_{2s}/R_{4s}/\ 3_{1s}/12_x$	N=3, R=4;
	N=6, R=2

Here is a list of some of the higher N multirules, which are for those seeking extreme quality control of extremely problem-prone methods. See the QC Application on Immunoassy QC with Higher N Multirules for more details on these rules. The power curves are also available online.

Higher N Multirules	
$1_{3s}/2\text{of}8_{2s}/4\text{of}8_{1s}$	N=8
$1_{3s}/3\text{of}8_{2s}/4\text{of}8_{1s}$	N=8
$1_{3s}/2\text{of}5_{2s}/3\text{of}5_{1s}$	N=5
$1_{3s}/2\text{of}6_{2s}/6_x$	N=6
$1_{3s}/2\text{of}8_{2s}$	N=8
$1_{3s}/3\text{of}8_{2s}$	N=8
$1_{3s}/2\text{of}6_{2s}$	N=6
$1_{3s}/2\text{of}5_{2s}$	N=5

When we say "known" we mean that the error detection and false rejection characteristics of the combination of rules are known. All this means is that probability studies have been performed. So for those rules listed in the table, we know how well they detect errors. For rules that aren't listed on the table, we have no idea. If you're using a rule not on the table, you're flying blind, crossing your fingers and just hoping that the rule actually detects the errors you need to detect.

We're not claiming that these are the only combinations that should be known. New combinations could be explored and probability studies could be performed. If the new rules display good error detection characteristics and low false rejection rates, then that's a new and valid addition to the rules listed above. In fact, email us if you know the characteristics of a new multirule.

12. Interpret multirules to help indicate the occurrence of random error or systematic error.

Why are the "Westgard Rules" the way they are? When we came up with them, did we pick all these rules out of a hat and stick them together? No, there was method to our madness. We chose each particular rule in the "Westgard Rules" because it was sensitive in a particular way to a particular kind of error. The error detection of the combination of those rules was in a way greater than the sum of its parts.

Quick review: there are two types of errors, random and systematic. Also coincidently, there are control rules which detect random errors better than systematic errors, and control rules that pick up systematic errors better than random errors. So the multirule combines the use of those two types of rules to help detect those two types of errors.

Here's a table listing the type of error and the control rule that best detects it.

Type of Error	**Control rule that detects it**
Random error	$1_{2.5s}$, 1_{3s}, $1_{3.5s}$ R_{4s}, $R_{0.05}$, $R_{0.01}$
Systematic error	2_{2s}, 4_{1s}, $2of3_{2s}$, 3_{1s} 6_x, 8_x, 9_x, 10_x, 12_x, $x_{0.05}$, $x_{0.01}$, cusum

Given this knowledge, when you get a particular control rule violation, you can begin to figure out what's wrong with the test method by looking at the rule that was violated. Was it a 1_{3s} violation or a R_{4s} violation? That's likely a random error. Were there 6, 8, 9, 10 or even 12 control results on one side of the mean? That's most likely a systematic error.

Resources:

QC - The Chances of Rejection

Conclusion

This is quite a list. I would say that most laboratories can't claim to have implemented all of these practices. Some laboratories may not be able to say they've implemented any of the points! **But these are *the* best practices**. If I knew a laboratory was doing all of these things with their testing, I would be extremely confident of the quality of their testing.

As I said earlier, many of these points are about QC best practices in general, not just "Westgard Rules" specific behavior. "Westgard Rules" are part of a QC context. Doing the best "Westgard Rules" means that you are doing the best all-around QC, too.

If you're reading this and find yourself at a loss of what to do, and/or where to start, fear not. Taking the first step is hard, but you quickly build up momentum. One quality improvement leads to another. The efficiencies and savings multiply as the quality improves.

If your QC happens to be in a poor state, there is still no reason to fear. That means there's a lot of room for improvement, and any step will make it better. Probably the best thing to do is not to try and tackle all the best practices, but instead try to eliminate all the worst practices first.

Finally, start small. Don't get overwhelmed by trying to change everything in the lab all at once. Try a pilot project first. Use the results to get management commitment from the powers above you. For those of you with highly automated chemistries, work on those first

ABUSES, MISUSES, AND IN-EXCUSES

WARNING! You may not want to read this article. It's a sobering list of all the common mistakes made by manufacturers and laboratories when they design, implement and interpret the "Westgard Rules." As it turns out, when your software or instrument or LIS claims to have "Westgard Rules," it might not be true or even useful. And if you see a claim that they've "modified" the rules to make them better, *be afraid*.

A Top 10 list of problems with QC and the "Westgard Rules"

- 10. Abuse of the term "Westgard Rules"
- 9. Misuse of "Westgard Rules" as a specific set of rules, namely $1_{2s}/2_{2s}/R_{4s}/4_{1s}/10_x$.
- 8. Misuse of the 12s "warning rule" in computer implementations.
- 7. "In-excuse" for using some inappropriate single-rules alone.
- 6. Misuse of the R_{4s} rule across runs.
- 5. "In-excuse" for illogical combinations of control rules.
- 4. Misuse of combinations of control rules whose error detection capabilities are not known.
- 3. "In-excuse" for not defining the details of rule implementation.
- 2. Misuse of "Westgard Rules" as a magic bullet.

- 1. Misuse of "Westgard Rules" when simpler QC will do.
- Conclusion

The "Westgard Rules" are now over 20 years old. Over the course of more than two decades, as you might imagine, we have gotten a lot of questions (and complaints and sometimes even curses) about the use and interpretation of "Westgard Rules." These questions come not only from those med techs working at the bench level, but from the manufacturers who want to include the "Westgard Rules" as an extra feature in their products. Many manufacturers claim to have implemented "Westgard Rules" in their instrument software, QC data workstations, and laboratory information systems. Unfortunately, many of those implementations just don't do it right. The result of poor implementation of the rules is often frustration, and the customers often blame those #$%&! "Westgard Rules" as the source of their troubles and problems.

To address the common questions and complaints, we've compiled a "Top 10" list of abuses, misuses, and "in-excuses" and other bad practices for the implementation of Westgard Rules. I warn readers that many of these points may hit close to home - in your own laboratories or your own instrument systems.

10. Abuse of the term "Westgard Rules."

If you read the original paper in the journal of Clinical Chemistry [CLIN CHEM 27/3, 493-501 (1981)], you'll find absolutely no use of the term "Westgard Rules." That term emerged from common usage of multirule QC procedures, probably as a shorthand way to identify the reference paper. I suppose it was too big a mouthful to say "multirule QC as described by Westgard, Hunt, Barry, and Groth." I'm not sure how this phenomenon got started, but it happened and now we're stuck with it. The problem is that there is no way to know exactly what someone means by "Westgard Rules." Many manufacturers claim to have implemented "Westgard Rules," but there's no way what they've done unless you test the performance of their QC software.

- Download and read the original Clinical Chemistry paper.

9. Misuse of "Westgard Rules" as a specific set of rules, namely $1_{2s}/2_{2s}/R_{4s}/4_{1s}/10_x$.

The original paper in the journal of Clinical Chemistry was intended as an example of the application of multirule QC, not a recommendation for a specific combination of control rules. The idea was to combine individual control rules to minimize false rejections and maximize error detection. Thus, we used (and still use) the broader term multirule QC to describe about this type of QC. Even in that paper, the need for adaptation of the rules was described based on the number of control measurements available.

- Read the "Westgard Rules" Homepage for a true definition of what the "Westgard Rules" are.

- Again, reading the original Clinical Chemistry paper will help.

8. Misuse of the 1_{2s} "warning rule" in computer implementations.

When multirule QC is implemented by a computer program, you don't need a warning rule. The reason for recommending a warning rule was that at that time - over 20 years ago - most QC was plotted by hand and interpreted visually. The 1_{2s} warning rule saved time in inspecting data manually; if there wasn't a 1_{2s} violation, you could skip those data points. With computer implementation, there is no need to start with a warning rule because the data inspection can be fast, complete, and effortless. The computer doesn't need to be "warned" - it has more than enough resources to check every point thoroughly.

- See the "Westgard Rules" Homepage for more about if and when to use 1_{2s}

7. "In-excuse" for using some inappropriate single-rules alone.

The 2_{2s} control rule by itself seems to be a favorite in some laboratories, if not by design then by default. Maybe the common practice of repeating controls when one exceeds a 2s limit has led to the routine use of a 2_{2s} rule by itself. However, this is really not a very good idea. That rule is responsive only to systematic error and it's not particularly sensitive by itself.

We use the term "in-excuse" because what's happening is that poor choice control rule is giving the laboratory an excuse to think that all the results are okay. The manufacturer allows the customer to choose a control rule that shouldn't be used by itself. The customer chooses that control rule out of some belief and possibly some experience that shows that there are fewer out-of-control flags and more in-control results when that particular control rule is used. Since the 2_{2s} single-rule undoubtedly has less false rejection that the 1_{2s} control rule, the method has fewer false rejects (which is good). In many cases, however, the control rule also isn't sensitive enough to detect a significant medical error that it should. So the control rule chosen doesn't sound many alarms at all, giving the customer the false sense of security that the QC must be great because problems are so rare.

Both the manufacturer and the customer are partly to blame. The customer is picking a control rule without real knowledge of how that control rule works. The manufacturer is allowing the customer to pick control rules that they shouldn't because the customer is always right. This co-dependency enables people to do bad QC. And it's inexcusable, since both the manufacturer and customer should know better.

- It might be worthwhile to review the QC - The Chances of Rejection lesson.

6. Misuse of the R_{4s} rule across runs.

The intention of the R_{4s} rule is to detect random error. When used across runs, systematic errors may be detected and misinterpreted as random errors. It is better to catch those systematic errors with the 2_{2s} or 4_{1s} rules to aid in trouble-shooting. Here again this is an "in-excuse" - a poor use of the control rule. Using the rules without any explanation or understanding of why the rules are combined the way they are. There's a deep logic to the combinations. Certain rules are good at detecting random errors, while others are good at detecting systematic errors.

- There's a good series of questions and answers about the R_{4s} here.

5. "In-excuse" for illogical combinations of control rules.

Multirule combinations should be built from the outside in. For example, when 2 control materials are analyzed, start with a single rule with wide limits such as 1_{3s}, then add a 2_{2s} and R_{4s}, followed by a 4_{1s}, and finally by a mean rule, such as 8_x, 10_x, or 12_x, depending on whether you want to "look-back" at the control data in the previous 3, 4, or 5 runs. When analyzing 3 control materials once per run, the rules fit better if you use 2of3_{2s}, 3_{1s}, and an appropriate mean rule, such as 6_x, 9_x, or 12_x to look-back at the previous 1, 2, or 3 runs. With 3 materials, it makes no sense to use an 8_x or 10_x control rule to look back at control results in the previous 1.7 or 2.3 runs.

4. Misuse of combinations of control rules whose error detection capabilities are not known.

Sure, you can combine any individual rules to make a multirule QC procedure, but only certain combinations have been studied and have known performance characteristics. Just because a computer program lets you pick and choose rules at random doesn't mean it's a good idea. Making up new combinations or rules is like making up new methods. There is a responsibility to document the performance of the new rules before you use them. This means performing mathematical calculations or simulation studies to determine the power curves and the probabilities for false rejection and error detection. Unless you can do that, you shouldn't make up new combinations of rules. The solution in QC software is to select from a list of defined multirule procedures whose power curves have been documented, rather than select individual rules to makeup the multirule procedure. Given a choice between a multirule combination whose performance is known and another whose performance is unknown, you should select the one that has documented performance characteristics.

- See some of the power functions for known combinations of "Westgard Rules"

3. "In-excuse" for not defining the details of rule implementation.

Multirule QC is actually simpler to do manually than by computer! The reason is that there are many possible rule applications within- and across-materials and within- and across-runs that must be explicitly defined in QC software. In manual applications, you can decide on the best or most appropriate way to inspect the data right when you're looking at the charts. In many software applications, it is not clear when control rules are being applied within- or across- runs, and/or within- or across-materials. And it is almost impossible to find a statement of how a particular software application implements the within/across rules.

- See the the article on Multirule Interpretation

2. Misuse of "Westgard Rules" as a magic bullet.

Just because you use "Westgard Rules" doesn't mean that you're doing the right QC. The most important detail about doing QC for a test doesn't concern the control rule used - the critical parameters are the quality required for the test and the bias and CV observed for the method. The control rule chosen flows directly from those details. And in some dire cases, when method performance is bad and CV and bias are high, no amount of "Westgard Rules" can help you. What you really need is a new method.

1. Misuse of Westgard Rules when simpler QC will do.

People are often surprised when we tell them that it may not be necessary to use multirule QC. You may not realize it, but in the labs were I work, not every test is QC'd with the "Westgard Rules." In fact, we only use "Westgard Rules" on those tests that are really hard to QC. Whenever possible, if a single control rule can provide desired error detection, then we'll do it that way because it's simpler and easier.

The across-the-board implementation of "Westgard Rules" on all instruments and all tests is not the most cost-effective way to manage the quality of the tests in your laboratory. It's important to optimize the QC for individual instruments and preferably for individual tests on those instruments. This can be done by following our Quality-Planning process that depends on the quality required for the test and the imprecision and inaccuracy observed for the method. You need to define the quality needed for each test to determine the right QC to implement.

Conclusion

So here you have a tally of "bad practices" that we've encountered when people, laboratories, and manufacturers implement the "Westgard Rules." I've always been a little afraid of pointing out the flaws in many implementations because I might discourage people from using multirule QC in their laboratory. If you've reached this paragraph, you may have the impression that "Westgard Rules" are so complicated you don't want to use them at all. But I want to assure you that it's really not that hard. Everything you need to know to be able to properly use the "Westgard Rules" is available for free (and on this website, in fact). Stay tuned to these pages for another article on "best practices" for "Westgard Rules" - as well as a way to tell if you're doing things correctly

TEN WAYS TO DO THE WRONG QC WRONG

Think you're the only one who doesn't do QC perfectly? You're not alone. In this article, we look at numerous examples from readers, where the best intentions have gone astray. By seeing how QC practices go wrong in the real world, we can learn what we need to do better.

Detecting (and avoiding) errors of the third kind

In an earlier essay, we described the difference between the “ease” of theoretical QC implementation and the complications of QC implementation in the “real world” laboratory. Outside the confines of a textbook or a website, laboratory professionals face multiple layers of problems when they approach QC; they not only have to interpret the QC rule correctly, they have to choose the correct QC rule, and they must base that rule choice on the correct data. No wonder then, that we often find laboratories not only doing the wrong QC, but doing the wrong QC wrong.

With this background and perspective, let’s look at some of those real world situations. We’re going to look at several of the out-of-the-blue, over-the-transom questions we receive at Westgard Web and dissect them into their component problems. Let’s see if and how they fit into the “wrong QC wrong” model and how we might help identifying the real problem.

Important Note: Anonymity has been preserved.

Names have been removed, unimportant details have been changed, and even the writing patterns have been altered to protect the identities of these questioners. We don’t want to discourage any feedback from customers and readers. We encourage you to contact us with your questions. The purpose of this essay is to examine typical real-world scenarios and discover common problems that all laboratories are facing.

Scenario #1:

“We are part of a laboratory network of approximately 20 labs, which use the same instruments and we are compared to one another. I use the Mean of the lab group then 2/3 of their SD's for my labs' Mean and SD's, and then adjust them through the first few months, as long as we stay in the groups' 2 SD. I would like your option of my use of the numbers.”

Here we see the problem with the use of the peer group statistics. Good information is being used for the wrong purpose. The peer group statistics should be used to compare performance between laboratories, for example, to determine bias – how each individual laboratory compares to the group, but shouldn’t be used to calculate the control limits for the individual laboratories. Each lab should use its own data to determine its own means and SDs for control limits.

The adjustment of the SDs, e.g. using 2/3 of the group SD, supports the idea that the group SD is too large for an individual laboratory, but use of an arbitrary factor to reduce the SD is just a guess. It’s best to get the means and SDs right for the individual laboratory. Otherwise, even if you “correctly” design QC and “correctly” interpret the control rules, you are still doing the wrong QC wrong. It’s like playing poker with alphabet flash cards.

Scenario #2:

“My lab runs 3 QC levels for TSH. The QC was very much within range but on day 5, the QC for level 2 was suddenly +3s, but QC 1 and QC 3 were within range. I accepted the run for that batch. On the next run, the QC for level 2 became -3s but the QC 1 and QC 3 were within range. The problem was later identified as deterioration of QC 2 material, and thus changed to a new QC 2 (on the next run all QC were within range). However, can I accept the second batch of run for patient's sample (for QC 2 -3S but QC 1 & QC 3 within range)?”

The QC design here is apparently to use 3 SD limits with 3 different control materials (1_{3s} with N=3 in our terminology). The correct interpretation for that control procedure is to reject a run if any one of the QC’s exceeds a 3s limit. It’s very unlikely that a control outside of 3s is a false rejection. Something is definitely wrong, which turned out to be the control material itself. That problem should have been identified on day 5 rather than waiting another day to see the problem again. Sounds like the same vials of controls are being used for several days, which assumes the controls are properly stored, remain stable, and do not get contaminated in any way.

Scenario #3

“1) We routinely accept assays with more than one QC out 2SD, if it is out on the side of positive screens. For instance, if we are screening for elevated phenylalanine and 2 or more QC are out +2SD, the run would be accepted and any patients that were elevated would be retested on the next assay.

“2) It had been standard practice that if one QC was out +2SD and one -2SD, they ‘evened out’ and the run could be accepted. (I see on the Westgard worksheets that this would violate the R4S rule and the run should be rejected.)

“Given that we are a screening lab as opposed to diagnostic, are these reasonable practices? CLIA took issue that we did not have published material stating that this was acceptable.”

It's harder to give advice about this situation because of the use of 2SD control limits, without knowledge of how the control limits are being calculated and whether they really represent 2SD of the method. If it is assumed that the control limits are correct for the method, any situation where 2 or more controls exceed their 2SD limits should identify real analytical problems, either systematic errors (2 out the same direction) or random errors (out different directions). In the case of both out high, the patients get retested, which is good. In the case of one out high and one out low, the errors may even out for the laboratory, but they don't for the individual patients. Some positive patients are likely misclassified as negative other some negatives are misclassified as positive. Any positives are retested by a confirmatory laboratory, which should sort out the false positives. Unfortunately, there may be some false negatives that don't get retested and that is a serious issue since the purpose of screening is to detect all potential positives. A screening lab should err on the side of false positives and avoid any false negatives.

To assure the right QC rules and right number of controls are being used, it is possible to apply our QC design process, even for qualitative tests. There is a detailed example for phenylalanine in Chapter 14 of Six Sigma Quality Design and Control. Like many labs, this laboratory appears to be applying blanket rules for QC, which means that individual methods may be over-controlled or under-controlled.

The immediate answer is, if you have decided to use the 2s control rule or the R4s control rule, you need to enforce it every time, not just selectively. And the R4s rule does not have an "even out" clause.

Secnario #4

"If I understand you correctly, we do nothing when the one qc is out and not even run it again. My friend says he runs it again to make sure it is random error.

" We had a problem with direct bili's and i must say that there were 2 levels out. It seems to me that when we do have problems that require hotline, there are 2 and sometimes all 3 levels out. I feel our inst may not have problems when only 1 is out. I will go back to the maintenance log of our analyzer and see if that has been the case.

"Regarding the practice of accepting some of these outlier (but less than 3sd) results, our ranges will be wider. However if we don't accept, the ranges will be tighter. I wonder if the ranges will stablize better and we will see the results fall above and below the mean if we did start accepting the outliers."

I don't think we've ever been on record as saying "do nothing when one qc is out." We do recommend that the only proper application of 2 SD limits is when only one control material is being run, i.e., N=1 and false rejections will be 5%. Other people may recommend doing nothing when one qc is out and that may become a common practice in some laboratories as a consequence of having a high level of false rejections.

The third paragraph is a perfect example of wrong QC wrong: by using the wrong limits, the laboratory also adopts the wrong practice for calculating control limits. The fundamental principle of statistical QC is to characterize performance under stable operating conditions in order to identify changes due to unstable operation, i.e., when problems occur. Thus, in principle, the control data from any out of control run should not be included in the calculation of control limits because it does not represent stable operation. Even the College of American Pathologists has this problem because of trying to cope with the use of wrong control limits, thus this practice is actually widespread.

Here's a case where proper design of the QC procedure is important so that any "out-of-control" signals represent real problems, not false rejections. If the control rules being used are tighter than necessary (for example, 2s limits), than the "outliers" may actually be acceptable if the QC were properly designed. And if the QC were properly designed, those outliers would actually be "in-liers" and they would rightly be used in the calculation of ranges. But a well-designed QC procedure that is properly implemented should not include data from "out-of-control" runs to calculate control limits. The problems of this scenario create a tortured logic that makes doing the wrong QC wrong seem like it's right.

Scenario #5

"I have fixed the lab limits based on the total allowable error, taking one fourth the value of the total allowable error for the analyte as 1SD. I have taken the target value from the assay sheet as my lab mean. Is my approach correct?"

Again, a little knowledge can be a dangerous thing. Using total allowable errors for quality requirements is good, but dividing them by four to generate a standard deviation figure is not correct. The standard deviation should be a measure of actual performance; total allowable error is a goal for the desired performance. You use your actual observed standard deviation to see if you are meeting the goal.

In this scenario, we also see that the target value from the assay sheet has been used for the laboratory mean. This may be acceptable when the control lot is first being introduced, but as soon as real data is available (from a short replication study, for example), that should be used instead of the assay sheet data.

Scenario #6:

"During daily practice in Clinical Chemistry lab if we get the 2 2s and after troubleshoot (recheck the analyser condition, the calibration, the reagent and the QC), we still get the same result as 2s . What should we do? If we still proceed doing the troubleshooting, I afraid the Turn Around Time of the urgent test will be longer than 45 minutes. This will effect the treatment of the patients."

This situation is a bit ambiguous, but let's assume that the "2 2s" means the 2_{2s} rule, rather than a 1_{2s} rule that was violated and the control then repeated. A 2_{2s} mean there is a real problem, but evidently the trouble-shooting didn't resolve it. The next QC was also out by 2s, in this case, most likely indicating that the problem hasn't been fixed.

The other issue here is a very common one for laboratories: production pressure. The TAT on tests is probably the most measured element of performance – and the element that is most recognized and felt by the ordering physicians.

Here's the real question: do you want to report possibly erroneous values to your physician just to make the TAT? Would it be acceptable to allow the doctor to make the wrong medical decision based on a wrong test result? Or do you want to get the values right the first time?

In a wrong QC wrong world, getting the numbers out the door becomes the highest priority, regardless of the quality of those numbers. That's like fast food testing – it's quick but it's bad for you. In a right QC right world, you'll have fewer alarms and out-of-control situations, so when you do get an error flag, you'll know it's a serious error and one that could affect patient values and treatment. You'll know that you don't want those values out the door.

Scenario #7

"In our laboratory value of many haematological parameters analysed few values are lying on one side of the mean in LJ graphs with no deviation. All the values are lying at one level with no deviation i.e more than 10. Our instrument is calibrated correctly. The parameters which are showing such pattern are MCHC, RDW. Kindly let me know the reason for this."

This laboratory has noted that the control points don't seem to show the expected distribution about the mean. There is possibly a very subtle issue here – data rounding. If precision is very good and the control results are being rounded, that could cause many points to show as the same value. It is often good practice to carry one extra significant figure in your control data so that the calculation of the mean and SD will be more accurate.

Another issue is using the right QC rules based on the quality required for the tests. It's quite likely that the 10:mean rule isn't needed if the control data are showing high precision. Getting the right calculations for the mean and SD are important, then getting the right QC design is possible. It looks like there are issues with the right QC and also with implementing QC right.

Scenario #8

"I have a case, my control measurement not exceeds a 2s control limit but 4 consecutive control measurement exceed the same mean plus 1s or the same mean minus 1s control limit (41s) or and 10 consecutive control measurement fall on one side of the mean (10x). What does it mean, accept run or reject run ? (12s (No) but 41s (yes) or 10x (Yes).? Accept run or reject run ?)"

This is actually a very straightforward "Westgard Rules" question. Nevertheless, we're going to make it complicated.

In the "classic" version of the "Westgard Rules," you only triggered the other rules after a 1_{2s} control rule was violated. So, strictly according to the classic rules, if there wasn't a 1_{2s} violation, then you don't use the other rules and everything is in.

Now, in the updated "Westgard Rules", the 1_{2s} "warning rule" has been replaced by a 13s rejection rule – and all the rules are to be used as rejection rules. If you were using the updated rules, that 4_{1s} or 10_x would be a rejection signal – but only if QC Design mandated that those 4_{1s} and/or 10_x mean rules were necessary to monitor the test. It's possible that the performance of this method, coupled to the quality required by the test, may make such rules unnecessary. That is, you might only need simple rules like 1_{3s} and can totally ignore those other data patterns.

Many laboratory professionals like to use the 10_x and 4_{1s} and similar rules as "warning rules," using those trends and shifts as a way to get an early eye on a problem, even if QC design doesn't mandate those rules. That's fine, but if it starts to make you chase ghosts in the method, it's counter-productive.

Scenario #9:

"I have a question regarding the 10x rule. If there are assays that consistently run above the established mean, but remain with in 2SD, does the run have to be rejected? Does the control have to be repeated? Can I legitimately adjust the mean to reflect how the control is performing?

"For instance: If our mean for potassium is set at 6.0 and our values consistently run 6.1, all of the values will fall on the same side of the mean. It seems unreasonable that the run should be rejected."

Believe it or not, this is the same problem as experienced by the laboratory professional in the previous scenario (Did you detect the similarities?).

Under the "classic" version of "Westgard Rules," there are no out-of-control values in the scenario. With the updated "Westgard Rules," we would in fact use those 10x rules and declare that run "out" – if in fact the 10_x mean rule was necessary. However, if the mean was adjusted to reflect current performance, a QC Design process might determine that none of the extended multirules were necessary. If the test is performing with barely any variation at all, it's more likely that a simple single rule, perhaps a 1_{3s}, will be chosen as the right QC procedure. Then those 10_x "violations" wouldn't count again.

Perhaps it's important to note this: just because a rule exists doesn't mean it needs to be implemented. It's tempting to lock on 10_x violations once you know about the 10_x rule. But there are times to use the rule and there are times to not use the rule.

Scenario #10

"If I have 6 analyzers performing Glucose, I run my controls 10-15 times on each analyzer to establish my mean and SD on each analyzer. Then I pool all of them and establish a grand mean across the board with CV. When I use this it is too narrow, because not all analyzers behave the same way to the exact precision. In this case how much I can loosen up the SD or CV, so that I can monitor the QC without causing too much problems? Is there a guideline or a standard I can use? Or what should be the criteria. Is there a maximum CV analytical I can use across the board?"

First, this is one example where we can be more specific because we're working with a single analyte: glucose. We know, for example, that the CLIA criteria for proficiency testing for glucose is 10%. That gives us an analytical goal for our QC Design process.

This laboratory is trying to do the right thing with their data. Pooling the data together to establish a grand mean can be useful – if it's used to assess bias between the analyzers. But calculating a peer sd from that data and applying it to all the individual instruments is not a good practice. In general peer sds are wider than individual instrument sds.

What’s surprising is that the laboratory has found this wider sd is still too narrow. Here is where we have to start making guesses. Let’s assume the “narrow” complaint actually stems from the use of 2s control limits. The number of false rejections caused by 2s limits is making the director conclude that the limits are too tight and should be loosened up.

Again, the end result of all this diligent effort and good intention is the wrong QC wrong. What actually needs to be done is that each instrument needs its own performance and its own QC chart and its own QC Design. That QC Design might actually result in those desired looser rules, and at the very least, the QC Design process eliminates 2s control limits. The pooled data can still be used to assess bias, which will be included in the QC Design process. While it may seem daunting to create specific rules for each instrument, it’s quite likely that they will all have very similar performance and end up with the same control rules anyway. But if one of those instruments is the runt of the litter, it better get different treatment from the lab.

Conclusion: It's easy to get it wrong

There may only be 50 ways to leave your lover, but this list of just ten problems shows us that there are probably an infinite number of ways to do the wrong QC wrong.

Let us stress this one more time: the people who submitted these questions and problems were not trying to get it wrong. They were well intentioned professionals trying to do things correctly. Believe us when we say there are far worse practices out there: laboratories that don't care about quality at all and would never consider asking a question about what's the right thing to do.

Doing the right QC right is not easy. One mistep and your good intentions can be led astray. That's why quality is so valuable - because it means you've taken the care at every step. That's why doing the right things right is so important - because it means you're delivering the best possible results to the physician and patient.

FAQ'S ABOUT MULTIRULE QC

Frequently-Asked-Questions (FAQs) about "Westgard Rules" and multirules.
Plus, some questions about Immunassays and QC (scroll down past the first section).

- When should you use the 4_{1s} and 10_x rules for "warning" rules and when should you use them as out-of-control rules?
- Should I have an 1_{2s} rule violation for starting evaluation of violations of 4_{1s}, 10_x, 8_x and 12_x rules?
- When would I use 8_x and 12_x rules?
- What is N?
- What's the best way to chart QC for multirule applications?
- Does the 1_{2s} warning rule have to be used in a computerized implementation?
- Can other rules be used as warning rules rather than rejection rules?
- Other than better error detection, are there any reasons to multirule procedures instead of single rules?
- What rules are most sensitive for detecting systematic errors?
- What causes systematic errors?
- What rules are most sensitive for detecting random error?
- What causes random errors?
- When can a single rule QC procedure be used instead of a multirule procedure?
- How do you decide if you need to apply rules across runs?
- When one rule in a multirule combination is violated, do you exclude just that control value from the QC statistics

New Questions about Multirule QC

1. In your article about multi-rules published in 1981 and in your book *Cost-Effective Quality Control: Managing the Quality and Productivity of Analytical Process* , page 95, you say that violation of the rules 4_{1s} and 10_x are signals of out-of-control and they lead to rejection of the run. In the same book, paragraph: modifications to use for warning purposes, pages 113 and 114, you say that those rules should be used as warnings for preventive maintenance in order to reduce false rejections. In your page on Internet in the link: "Multirule and 'Westgard' rules: what are they?" the rules are again signals of out-of-control runs. Could you clear this subject to me?

One situation might be methods on instrument systems where periodic changes in reagents introduce small systematic errors that can't be easily or completely eliminated by recalibration. These systematic changes may be judged to be medically unimportant, but the 4_{1s} and 10_x rules might still detect them. In such a case, it may be useful to apply the 4_{1s} and 10_x rules when you change the lot number of reagents and want to check for small shifts, but then "turn-off" those rules once you decide there aren't any shifts or when an observed shift is judged to be small and not important to detect. Otherwise, those rules will continue to give rejection signals even though you have decided not to do anything about the problem.

You can make these decisions on what rules to apply much more objectively if you follow our recommended QC Planning Process, where you define the quality required for the test, account for the imprecision and inaccuracy observed for your method, then select the control rules and numbers of control measurements necessary to detected medically important errors.

2. Should I have an 1_{2s} rule violation for starting evaluation of violations of 4_{1s}, 10_x, 8_x and 12_x rules?

In the original multirule paper, we recommended the use of the 1_{2s} rule as a warning, then the inspection of the data by the other rules to decide whether or not the run should be rejected. This was done to simplify the manual application of the rules and keep from wasting time when there wasn't likely to be a problem. While in principle it is possible that there might be a 4_{1s} or 10_x violation without ever exceeding the 2s warning limit, our experience has been that it seldom happens, at least if the control charts have been properly set up and the limits calculated from good estimates of the observed method variation.

In computer assisted applications of the multirule procedure, there is no need to use a 2s warning limit. All the chosen rejection rules can be applied simultaneously.

3. When would I use 8_x and 12_x rules? What are the advantages of these rules over 4_{1s} and 10_x in the indication of systematic errors?

You pick the number of consecutive control measurements to fit the product N times R, where N is the number of control observations in the run and R is the number of runs. For example, if N is 4 and R is 2, you would want to use the 8_x rule to look at the control data from the current and previous run, or a 12_x rule to look back over three consecutive runs. A 10_x rule would require looking back 2.5 runs, which doesn't make any sense. You would either look at the control measurements in 2 runs or in 3 runs, not 2.5 runs.

What is N?

When N is 2, that can mean 2 measurements on one control material or 1 measurement on each of two different control materials. When N is 3, the application would generally involved 1 measurement on each of three different control materials. When N is 4, that could mean 2 measurements on each of two different control materials, or 4 measurements on one material, or 1 measurement on each of four materials.

In general, N represents the total number of control measurements that are available at the time a decision on control status is to be made.

What's the best way to chart QC for multirule applications?

You can chart your QC data using regular Levey-Jenning's control charts, on which additional lines have been drawn to represent the mean plus/minus 1s, plus/minus 2s, and plus/minus 3s. You can set up one control chart for each level of control material being analyzed. These individual charts have the advantage of showing your method performance at each control level. However, it is difficult to visually combine the measurements from consecutive control measurements on different materials.

To combine measurements on different materials, you can first calculate the difference of each control observation from its expected mean, divide by the expected standard deviation to give a z-score or a standard deviation index (SDI), and then plot the SDI value on a control chart whose central mean is zero and whose control limits are drawn as plus/minus 1, plus/minus 2, and plus/minus 3. You can plot the values for the different materials using different colors to help keep track of trends within a material. This is a lot of work if you have to do it by hand, but many computerized QC programs support this type of calculation and often provide an SDI chart.

Does the 1_{2s} warning rule have to be used in a computerized implementation?

No, it was mainly intended for manual implementation to trigger the application of the other rules. When you apply the rules manually, it sometimes is a lot of work to look through all the control data and check it with several rules. If the computer can do all the rule checking, then it's not much work for the analyst to apply all the rules and there's really no need to apply the 1_{2s} rule at all.

Can other rules be used as warning rules rather than rejection rules?

There's another type of warning rule that can be used to indicate prospective action instead of run rejection. With very stable analytical systems, it may be advantageous to interpret rules like the 1_{4s} and 10_x as warning rules because they are quite sensitive to small shifts that occur from run to run, day to day, or reagent lot to reagent lot. If you do this, you also need to define the response that is appropriate for a warning. That may be to perform maintenance before the next run, carefully inspect the analytical system, review system changes, review patient data, etc.

Other than better error detection, are there reasons to use multi-rule procedures instead of single rules?

If the same error detection is available by multirule and single rule QC procedures, but that error detection is less than 90%, then it would still be valuable to use a multirule procedure because it can increase the number of control measurements inspected by applying rules across runs, thereby improving the detection of persistent errors (i.e., errors that begin in one run and persist until detected).

Another potential advantage of a multirule procedure is that the rule violated can provide a clue about the type of analytical error that is occurring. For example, violations of rules such as 2_{2s}, 4_{1s}, and 8_x are more likely due to systematic errors, whereas violations of rules such as 1_{3s} and R_{4s} are likely due to random errors.

What rules are most sensitive for detecting systematic errors?

Rules like the 2_{2s}, 3_{1s}, 4_{1s}, 6_x, 8_x, 9_x, 10_x, and 12_x tend to be more sensitive to systematic error than random error.

What causes systematic errors?

Systematic errors may be caused by inaccurate standards, poor calibration, inadequate blanks, improperly prepared reagents, degradation of reagents, drift of detectors, degradation of instrument components, improper setting of temperature baths, etc.

What rules are most sensitive for detecting random error?

Rules like the $1_{2.5s}$, 1_{3s}, $1_{3.5s}$, and R_{4s} are most likely to detect random errors.

What causes random errors?

With automated systems, random errors may be due to incomplete mixing, bubbles or particles in the reagents, probe and syringe variations, optical problems, sample line problems, etc. With multitest use of a control material, apparent control problems on several tests may actually be a random or individual problem with the control material itself. With manual methods, random errors may be due to alliquoting and pipetting, timing variations in critical steps, readout variation from cell to cell, etc.

When can a single rule QC procedure be used instead of a multirule procedure?

In the case where a single rule procedure provides 90% detection of the critical-sized errors in a single run, any problems that occur will generally be detected right away, in the first run in which they occur. The single rule procedure may be simpler to implement and therefore be preferable for such applications. With high precision automated chemistry and hematology analyzers, there may be many tests for which a single rule QC procedure is perfectly adequate.

How do you decide if you need to apply rules across runs?

You need to first assess the error detection within the run by use of a critical-error graph or an OPSpecs chart. If error detection is less than 90%, then it will generally be advantageous to apply rules across runs to detect persistent errors as soon as possible.

When one rule in a multirule combination is violated, do you exclude just that control value from the QC statistics?

No, you exclude all the control values in that run. Remember the QC statistics are supposed to provide estimates of the mean and SD to represent stable method performance, which is then used to compare with current performance. All control results in an out-of-control run are suspect of not representing stable performance.

Immunoassay QC

This month's second question comes from Brisbane, Australia:

A website visitor raised some issues about QC for immunoassay methods. He noted that the daily controls are well within the manufacturer's specifications - so much so they often form a straight line over periods of time and are not randomly distributed about the mean as one would expect.

- Are the manufacturer's specifications for acceptable control values too wide?
- Should we set our own control limits based on our control data?
- How do you use control charts on extremely stable immunoassay analyzers?
- How do you determine the frequency with which to run controls on extremely stable analyzers?
- Where can I find some example QC planning applications for immunoassay methods?

Are the manufacturer's specifications for acceptable control values too wide?

Manufacturers sometimes use controls to ascertain whether their systems are behaving in a normal manner or whether they need to troubleshoot the system. In doing so, they may set the limits of acceptable values to encompass the performance expected from most of their systems in the field. These limits may reflect both within laboratory variation and between laboratory variation, and, therefore, may be much wider than appropriate for QC within a single laboratory.

Should we set our own control limits based on our control data?

Yes, this is the best practice for achieving tight control of a testing process. This practice allows you to optimize the testing process in your laboratory for cost-effective operation. You can then take into account the quality required in your laboratory when you select control rules and numbers of control measurements. You can follow our QC planning process and apply our QC planning tools to help you set appropriate control limits.

How do you use control charts on extremely stable immunoassay analyzers?

Make sure you assess stability relative to the critical sizes of error that would be medically important in your laboratory, rather than the manufacturer's specifications. If the method operates well within the quality required in your laboratory, you will be able to employ single rules such as $1_{3.5s}$ or 1_{3s} with a low number of control measurements. These QC procedures will assure a low rate of false rejections, and therefore contribute to cost-effective operation of your testing processes.

How do you determine the frequency with which to run controls on extremely stable analyzers?

This is a tough question and we don't have an easy answer!

In principle, if the analyzer were perfectly stable and never had a problem, it wouldn't be necessary to run any controls. On the other hand, if the analyzer frequently has problems, then it is necessary to run controls very often. Deciding how often to run controls is difficult because we seldom have adequate information about the stability of an analyzer under the operating conditions of our own laboratory. Furthermore, it is difficult to transfer any information from other laboratories because they may operate the analyzer under conditions that are different from ours.

The frequency of problems could be assessed if we were able to implement an ideal QC design that detects all medically important errors and gives essentially no false rejections. Then we could count the number of rejections, compare to the total number of acceptances plus rejections (or total number of runs) to determine the rate or frequency of problems, then optimize the QC design based on the frequency of problems.

Alternatively, we can make some judgment on the factors which make an analyzer susceptible to problems, such as a change in operator, change in reagents, recalibration, maintainence, etc., then establish the appropriate times for checking the operation of the analyzer. Manufacturers will generally recommend the maximum length of time that the analyzer can be run without rechecking performance. Laboratories may need to set a shorter length of time based on their operating conditions.

One novel way of measuring run length would be to use patient data to monitor the stabilitiy of the analyzer, as we discussed in a recent article in Clinical Chemistry:

- JO Westgard, FA Smith, PJ Mountain, S Boss. Design and assessment of average of normals (AON) patient data algorithms to maximize run lengths for automatic process control. Clin Chem 1996;42:1683-1688.

The ability to implement this approach will depend on the workload of the laboratory. The approach is probably most workable in high volume automated laboratories.

So, in practice we usually end up making our best judgment of when to run controls on the basis of the maximum period allowable according to the manufacturer's recommendations, our knowledge of the methods and their susceptibility to problems, our experience with how often we have problems with the analyzer, and the factors that affect operation of the analyzers in our own laboratories.

Where can I find some example QC planning applications for immunoassay methods?

We provided some detailed examples for prolactin, total b-hCG, CEA, FSH, LH, TSH, and b2-microglobulin in a recent paper:

- K Mugan, IH Carlson, JO Westgard. Planning QC procedures for immunoassays. Journal of Clinical Immunoassay 1994;17:216-222.

The application of higher N multirule procedures to immunoassay methods was discussed in a recent continuing education publication:

- JO Westgard, CA Haberzettl. Quality control for immunoassays. AACC Diagnostic Endocrinology and Metabolism: An in-service training & continuing education program. 1996;14(September):239-243.

Neill Carey has written a couple of good application papers:

- RN Carey. Quality- control rules for immunoassay. AACC Endocrinology and Metabolism: In-service training and continuing education 1992;10(September):9-14.
- Carey RN, Tyvoll JL, Plaut DS, Hancock MS, Barry PL, Westgard JO. Performance characteristics of some statistical quality control rules for radioimmunoassay. J. Clin Immunoassay 1985;8:245-252.

We also include example applications for cortisol, thyroxine, and FSH in our OPSpecs Manual - Expanded Edition (pp 5-30 to 5-35).

Check the archives for more Questions

The particular recommendation (on pages 113-114 of Cost-Effective QC) concerned when you might choose not to use the 4_{1s} and 10_x rules or when you might apply them as "warning" rules that are used prospectively to trigger inspection and preventive maintenance rather than apply them to reject a run

MULTIRULES AND QC VALIDATOR

A discussion of how multirule procedures are implemented in the QC Validator program. These answers also apply to the latest version of our QC Design software, EZ Rules 3.

Also included is a discussion of what a "run" is and how to define it for today's modern random access analyzers. Plus patient averages and moving averages. (Scroll down past the first section)

Analytical runs, placement of controls, patient averages, and moving averages

Application of multirules in the QC Validator program

- How are the different levels of control incorporated in the power curves?
- How does the number of control materials affect automatic QC selection?
- How does the number of runs (R) affect the selection of control rules?
- What are the implications of R in the simulation of power curves and the application of rules in a multirule procedure?
- What rules are used across runs?
- Why is the R_{4s} rule restricted to a single run?
- What about using the R_{4s} across materials within a run?
- Does the R_{4s} rule require consecutive control measurements?
- Is the R_{4s} rule violated if one control is +2.4 SD and another is _1.8 SD?

A user of the QC Validator program raised some interesting questions about how the control rules for multirule procedures are simulated and applied.

How are the different levels of control incorporated in the power curves?

For simplicity, we have made assumptions that method performance and analytical errors are the same at each level or each control material, so all control measurements will provide the same information about what's happening. In a sense, the performance and errors are normalized relative to the mean of the control material and considered much like an SDI type of value. Thus, the control rules can then be applied as if a single level of control material is being used and the level or material itself does not matter.

If these simplifying assumptions are not made, the process becomes much more complicated and requires that the power curves be determined for each specific set of conditions. Using a QC simulation program on a microcomputer, this might take several minutes for each power curve. Given the variety of rules and Ns of interest, it could take several hours or even a day to generate all the information needed to select appropriate control rules. Those conditions would vary to some extent from test to test, thus making the job of selecting QC procedures too laborious to be practical.

How does the number of control materials affect automatic QC selection?

The control rules considered during automatic QC selection depend on the total number of control measurements available, which in turn depend on the number of control materials. The number of control materials sets the minimum number of control measurements to be considered, but multiples of that number are also considered, usually corresponding to measurement each control material twice. For example, for 2 Materials, the total Ns considered by the default settings for the automatic QC selection criteria are 2 and 4; for 3 Materials, the total Ns are 3 and 6. Users can modify the selection criteria to eliminate the higher total Ns.

As an example, with automatic selection for 3 Materials, the autoselect feature will consider the 2of3_{2s} and 3_{1s} rules rather than the 2_{2s} and 4_{1s} rules. With a total N of 6, the autoselect feature will consider a 6_x rule rather than a 10_x rule.

How does the number of runs (R) affect the selection of control rules?

R here refers to the number of runs in which the control rules are applied (don't confuse this with the R in R_{4s} which refers to a range control rule). Most control rules are applied within a single run, thus R usually is 1. However, with multirule procedures, some rules cannot be applied within a single run if the total N is too small. For example, with a $1_{3s}/2_{2s}/R_{4s}/4_{1s}$ multirule procedure, only the 1_{3s}, 2_{2s}, and R_{4s} rules can be applied in a single run (R=1) having 2 control measurements. If R were 2, then the 4_{1s} rule could be applied by combining the 2 control measurements from the current with the 2 control measurements from the previous run, hence, using the 4_{1s} rule to look back at earlier control data and improve the detection of systematic errors that persist from run to run until detected and eliminated.

The default setting for R for the automatic selection criteria is always 1, regardless whether the number of control materials are 1, 2, or 3. This means that multirule procedures will be selected only on their capabilities for detecting errors in the first run. However, in situations where ideal error detection cannot be achieved in the first run, users could change the setting for R with the objective of selecting a multirule procedure that would achieve the desired error detection with a set number of runs.

What are the implications of R in the simulation of power curves and the application of rules in a multirule procedure?

In generating power curves, the control rules that can be applied within a run are always used in the current run (R=1) if the total N is sufficent. If there is a rule that cannot be applied in the current run, then if R>1, that rule will be applied if RxN is sufficent. However, the rules that were applied in R=1 are not applied again because they would have already be applied to the earlier run.

In applying multirule procedures to laboratory testing processes, the rules that can be used within the current run should be applied first to decide the control status of the current run, then any rules that require more than one run are applied next to detect persistent errors that cannot be detected in a single run.

For example, if a 1_{3s}/2of$3_{2s}/3_{1s}/6_x$ multirule procedure were applied for N=3 and R=2, the 1_{3s}, 2of3_{2s}, and 3_{1s} would be applied to the 3 control observations in a single run and the 6_x rule would be applied across runs to consider the 3 previous control observations as well as the 3 current observations. If the control observations in the previous run were +0.5SD, +2.2SD, and +1.7SD, and those in the current run are +2.1SD, 1.8SD, and 1.7SD, in that order, then a 6_x rule violation has occurred. Note, however, that because 2of3_{2s} rule can be applied within the run, there is no violation of that rule, even though there there is a sequence of +2.2SD, +1.7SD, and +2.1SD that occurs across runs.

What rules are used across runs?

Rules such as 4_{1s}, 6_x, 8_x, and 10_x are often applied across runs to detect persistent systematic errors. The 2_{2s} and 3_{1s} could also be used within a material and across runs if it is of interest to monitor systematic changes at one level, such as the high or low end of the working or linear range.

Why isn't the R_{4s} rule used across runs?

Remember that the R_{4s} rule is intended to detect random error, whereas the 2_{2s} rule is aimed at systematic error. If a systematic change occurred between two runs, perhaps due to a calibration problem, the R_{4s} rule would respond to this systematic change as well as any change in random error. Because we want to use these rules to give us some indication of the type of error occurring (which would help us in trouble-shooting the method), we prefer to use the R_{4s} rule only within a run and detect between-run systematic changes with other rules such as the 2_{2s} or 4_{1s}. Not everyone agrees with this and some analysts choose to apply the R_{4s} rule across runs, so be sure to look carefully at how the R_{4s} rule is being applied in your own laboratory.

What about using the R_{4s} across materials within a run?

We do apply the R_{4s} rule across materials within a run, even though it could be argued that a systematic change might occur at one level and not at the other, thus in principle the rule should not be used across materials, particularly if control limits were being calculated on the basis of within run SDs rather than a more long-term SD that represents the performance expected over many runs. Again, some judgment is needed here and you need to carefully define how to apply the rule in your method protocols, or carefully read the laboratory procedure manual to understand its intended use in your laboratory.

Does the R_{4s} rule require consecutive control measurements?

No, the R_{4s} rule is used to consider the highest and lowest observations in a group of control measurements, thus there is no requirement for consecutive observations like with the 2_{2s} or 4_{1s} rules. "Consecutiveness" is helpful for observing a shift in the mean of a distribution, i.e., systematic errors, whereas random error is observed by looking for the width of distribution which is more easily observed from the range of a group of observations.

Is the R_{4s} rule violated if one control is +2.4SD and another is -1.8SD?

No and yes, depending on whether the rule is defined as a qualitative counting rule or a quantitative range rule.

The original application of the multirule procedure was to ***count*** the number of measurements exceeding certain limits, therefore, it is a counting type of algorithm. Does 1 measurement exceed a 3s limit, do 2 in a row exceed the same 2s limit, does 1 in a group exceed the +2s limit and another exceed the - 2s limit, do 4 in a row exceed the same 1s limit, and do 10 in a row fall on one side of the mean? If R_{4s} is used as a counting rule, observations of +2.6SD and -1.8SD do not represent an R_{4s} violation.

If you want to be more quantitative and actually ***calculate*** the difference between the highest and lowest observations, then it is possible to use a quantitative range rule such as $R_{0.05}$ or $R_{0.01}$, in which case an observed range of 4.4SD would be a violation if N were 2-4 per run. These rules are usually used together with mean rules, which is the original QC recommendation developed by Shewhart and still widely used in industry today. QC Validator contains power curves for mean/range procedures and the automatic selection criteria can be modified to select these procedures if they can be implemented in your laboratory.

Another user from the Netherlands provides a series of questions about how to define an analytical run, placement of controls within a run, and the use of patient data and moving average QC procedures.

- How is a "run" defined for automated random access analyzers?
- Does it make any difference whether a pair of control materials are analyzed immediately, one after the other, or in random order, separated by time, say one in the morning and one in the afternoon?
- Is it important to include patient averages to assure quality and detect preanalytical as well as analytical factors that may not be observed with control samples?
- How do QC procedures that make use of moving averages compare to multirule procedures?

How is a "run" defined for automated random access analyzers?

For instance, a run could be all patient samples between two controls, one full tray, or one shift. It makes quite a difference if results are considered validated only after the last controls have been analyzed.

We touched on a similar question earlier and acknowledged that "this is a tough question and we don't have an easy answer." That answer is still true, but it may be useful to discuss this a bit more.

The above question also implies that the definition of a run includes the practice of "bracketing" patient samples by controls. It is important to understand that the practice of bracketing dates back to early continuous flow analyzers that were not very stable and tended to drift over a rather short period, such as the time required to analyze a tray of samples. It became standard practice to place controls at the beginning of the run - right after the calibrators - and at the end of the samples or the end of the tray, whichever came first. If the controls at the end of the run were out, it was usually due to a problem of drift. The results on the patient samples had changed significantly from the beginning of the run, therefore, it was sound practice to repeat the samples in between the controls.

Today's fourth generation analyzers have quite different operating characteristics (see discussion of Future Directions in Quality Control), which suggests that the practice of "bracketing" a group of samples with controls may not be appropriate. Better guidance for defining a run is provided by NCCLS [Document C24-A. Internal Quality Control Testing: Principles and Definitions. National Committee for Clinical Laboratory Standards, 940 West Valley Road, Suite 1400, Wayne, PA 19087-1898], which provides the following definitions:

- "**Analytical run:** For purposes of quality control, an analytical run is an interval (i.e., period of time or series of measurements) within which the accuracy and precision of the measuring system is expected to be stable. Between analytical runs, events may occur causing the measurement process to be susceptible to variations that are important to detect.
- "**Manufacturer's Recommended Run Length (MRRL):** The manufacturer should recommend the period of time or series of measurements within which the accuracy and prrecision of the measuring system, including instruments and reagents, are expected to be stable.
- "**User's Defined Run Length (UDRL):** The user should define the period of time or series of measurements within which validation of the measurement process is important based on the stability, reporting intervals of patient results, cost of reanalysis, work flow patterns, operator characteristics, or similar nonanalytic considerations that are in addition to the expected stability of the accuracy and precision of the measuring system."

These statements suggest that the run be defined in units of time or units of samples based on the expected stability of the method, the size of changes that would be important to detect, and the changes in conditions that make the method susceptible to problems. While the maximum period of a run is defined by the manufacturer, the user is responsible to assess laboratory factors that may require a shorter run, thus definition of a run is a shared responsibility of the manufacturer and the user. Regulations sometimes set another maximum, such as CLIA's 24 hour period as the maximum run length. In addition, it should be recognized that manufacturer's seldom deal with the issue of what size of changes would be important to detect, thus the user is really left with the responsibility of defining both the quality requirement and the run length.

With today's high-stability, high-precision, random access analyzers, it often makes sense to define run length in units of time. It also is practical to analyze controls initially, before patient specimens, in order to assure the system is working properly before starting patient analyses, then to monitor periodically to check performance. This implies a multistage QC design, with high error detection during startup and low false rejections during monitoring.

With certain electrode type analyzers where exposure to specimens may in some way "build up" and cause problems, it may make sense to define the run length as a certain number of specimens.

Does it make any difference whether a pair of control materials are analyzed immediately, one after the other, or in random order, separated by time, say one in the morning and one in the afternoon?

In selecting control rules and numbers of control measurements, our QC planning approach is to determine what control rules and how many control measurements are necessary to assure that an out-of-control signal will be obtained if medically important errors are present. This means if N=2, those two measurements are needed to determine the control status of the method. If you wait till the afternoon to get the second measurement, you won't know if the method is working properly until then; meanwhile, you may have reported a lot of patient results. Again, with modern instrument systems, we would argue for a multistage QC procedure having a startup design that will assure the necessary quality is being achieved before analyzing any patient samples, then spacing controls over time to look for changes in performance. It makes sense that controls spaced out over the course of a run would provide the best potential for picking up a problem as early as possible.

Is it important to include patient averages to assure quality and detect preanalytical as well as analytical factors that may not be observed with control samples?

For tests where stable control materials are available, patient data QC procedures, such as Average of Normals (AON), usually provide a secondary and complementary method for monitoring method performance. They may be useful in picking up preanalytical problems that reflect improper processing and storage of specimens, as well as analytical problems that do not show up in the same way on control materials. In general, AON procedure are more complicated to design because additional factors need to be considered, such as the ratio of the population to analytical SDs and the truncation limits chosen [see Cembrowski GS, Chandler EP, Westgard JO. Assessment of 'Average of Normals' quality control procedures and guidelines for implementation. Am J Clin Pathol 1984;81:492-499]. They also tend to be more difficult to implement and are impractical in many laboratory situations because of the high N that is needed to provide the desired error detection.

However, power curves can be determined and then we can apply the same QC selection and design methodology using OPSpecs charts and/or critical-error graphs. We recently illustrated how to do this and recommended using AON as a way to measure run length in automated process control systems [see Westgard JO, Smith FA, Mountain PJ, Boss S. Design and assessment of average of normals (AON) patient data algorithms to maximize run lengths for automatic process control. Clin Chem 1996;42:1683-1688].

How do QC procedures that make use of moving averages compare to multirule procedures?

QC procedures that employ a moving averages would be expected to perform similarly to a traditional mean rule, which is expected to have at least as good error detection as multirule procedures and possibly even better. We provide power curves for traditional mean/range QC procedures in the QC Validator program, along with power curves for a variety of multirule procedures. Parvin has recommended a multimean type of QC procedure that should have better error detection than a traditional multirule procedure [see Parvin CA, Comparing the power of quality-control rules to detect persistent systematic error, Clin Chem 1992;38:356-363].

POWER FUNCTION GRAPHS OF MULTIRULES

See the power functions of the different controls rules and combinations of multirules that make up the "Westgard Rule"

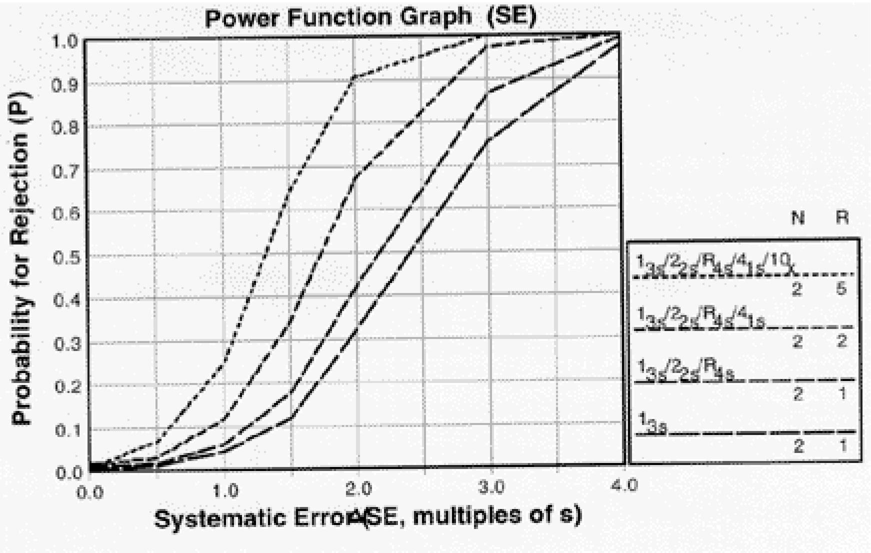
Power Function Graphs

Ns of 2 for systematic error
Ns of 2 for random error
Ns of 4 for systematic error
Ns of 4 for random error

The rationale for using multirule procedures and the expected performance characteristics of Westgard rules can be shown by power function graphs. See the lesson on Power Function Graphs for more background information and a better understanding of statistical power, power curves, and power function graphs.

Ns of 2 for systematic error

The accompanying figure shows a power function graph for systematic error, i.e., the probability of rejection on the y-axis versus the size of systematic error on the x-axis. Think of a 2s SE as a systematic shift equivalent to 2 times the SD of the method.

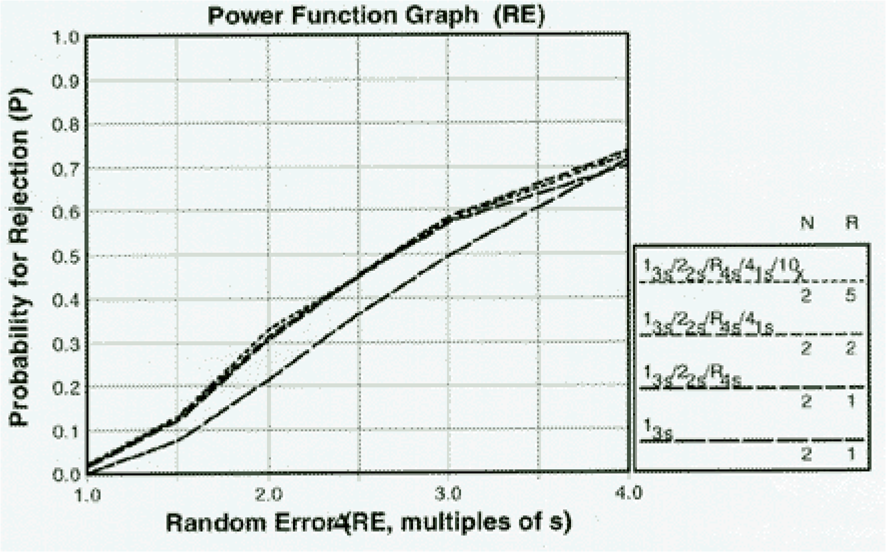


Power function graph for detection of systematic error showing effects of combinations of control rules with Ns of 2. Note that the order of lines in the graph match the order of lines shown in the key at right (i.e. the top line in the graph matches the top line of the key)

When N=2, the error detection available from a 1_{3s} control rule is shown by the bottom power curve. The increase in power from adding the 2_{2s} and R_{4s} rules is shown by the second curve from the bottom. Use of the 4_{1s} rule to "look-back" over two consecutive runs provides additional detection of persistent systematic errors, as shown by the next to the top curve (for N=2, R=2). This can be further increased when the 10_x is used to look-back over five consecutive runs (top curve, R=5).

Ns of 2 for random error

The power function for random error again shows the probability for rejection on the y-axis versus increases in random error on the x-axis. Think of an RE of 2.0 as a doubling of the SD of the method.

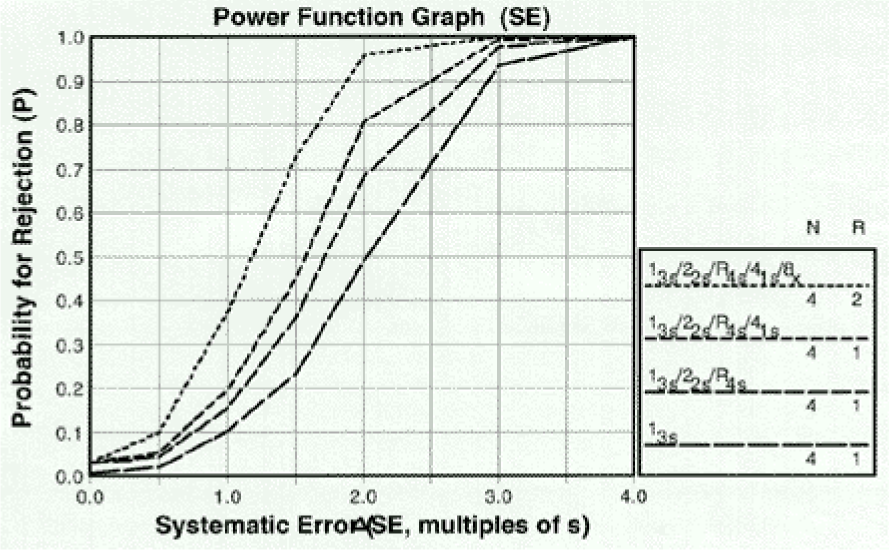


Power function graph for detection of random error showing effects of combinations of control rules with Ns of 2.

Random error is detected mainly by the 1_{3s} and R_{4s} rules. We have recommended that rules for random error, such as R_{4s}, be used only within a run in order to distinguish random error from between run systematic errors (which should be detected by rules such as 4_{1s}, 10_x). Therefore, the power of the multirule combinations show no further improvements in detection of random errors for across run applications; the power curves for the R-2 and R=5 combinations that use the 1_{3s} and R_{4s} rules essentially coincide with the power curve for those rules with R=1, as shown by the top three curves on this power function graph.

Ns of 4 for systematic error

Further improvements in error detection can be expected when N is increased to 4 because the 1_{3s}, 2_{2s}, R_{4s}, and 4_{1s} can now all be applied within the run.

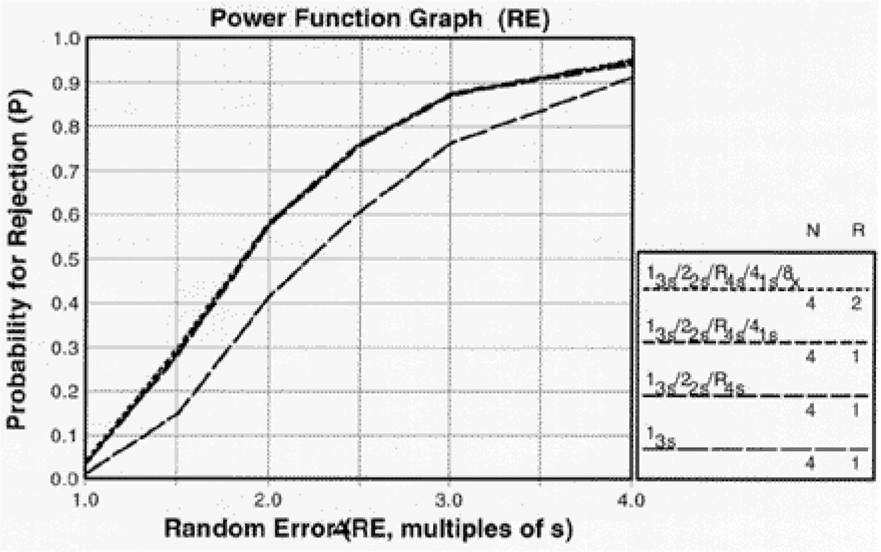


Power function graph for detection of systematic error showing effects of combinations of control rules with Ns of 4.

For detection of systematic errors, the original error detection of the 1_{3s} control rule is shown by the bottom power curve, the increase from addition of the 2_{2s} and R_{4s} rules is shown by the next to bottom curve, and the increase from use of the 4_{1s} rule within run is shown by the next to top curve. Error detection across runs is increased by the use of the 8_x rule to look-back over two consecutive runs (N=4, R=2).

Ns of 4 for random error

The detection of random error is again affected primarily by the 1_{3s} and R_{4s} rules. The original error detection from the 1_{3s} rule is shown by the bottom curve and is increased by the addition of the R_{4s} rule as shown by the top curves.



Power function graph for detection of random error showing effects of combinations of control rules with Ns of 4.

For more information, see these references:

- Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981;27:493-501.
- Westgard JO, Barry PL. Improving Quality Control by use of Multirule Control Procedures. Chapter 4 in Cost-Effective Quality Control: Managing the quality and productivity of analytical processes. AACC Press, Washington, DC, 1986, pp.92-117.
- Westgard JO, Klee GG. Quality Management. Chapter 16 in Fundamentals of Clinical Chemistry, 4th edition. Burtis C, ed., WB Saunders Company, Philadelphia, 1996, pp.211-223.
- Westgard JO, Klee GG. Quality Management. Chapter 17 in Textbook of Clinical Chemistry, 2nd edition. Burtis C, ed., WB Saunders Company, Philadelphia, 1994, pp.548-592.
- Cembrowski GS, Sullivan AM. Quality Control and Statistics, Chapter 4 in Clinical Chemistry: Principles, Procedures, Correlations, 3rd edition. Bishop ML, ed., Lippincott, Philadelphia, 1996, pp.61-96.
- Cembrowski GS, Carey RN. Quality Control Procedures. Chapter 4 in Laboratory Quality Management. ASCP Press, Chicago 1989, pp.59-79.