Quantitating Quality: Best Practices for Estimating the Sigma-Metric

Sten Westgard, M.S., Westgard QC

Abbott

Introduction

In a previous white paper, the theoretical basis for the Sigma-metric assessment of analytical quality was discussed. In this paper, we will discuss the optimal use of Sigma-metrics, from best practices for collecting the data used to calculate the Sigma-metric to the desirable implementation and operational implications of the Sigma-metric.

We will again use a simple analogy to enhance the understanding of the Sigma-metric equation: Sigma-metric = (TEa – Bias)/CV. Imagine a target with a bull's-eye surrounded by multiple rings, each of the rings designating a Six Sigma level of quality. The target is the quality requirement – the quality required by the test method in order to deliver the required analytical performance for proper clinical care, typically expressed as the Total Error Allowable (TEa). The test method is the arrow shot at the target, with the measured precision and bias providing information on performance. The Sigma-metric tells us whether the test method has hit the bull's-eye, or how close or how far away from the target the arrow (method) has landed.

Which target are we trying to hit?

There are three key components to the calculation of a Sigma- metric for a laboratory method: the quality requirement, the imprecision, and the bias. But there is a fourth critical component that goes unstated: where in the reportable range (or dynamic range) of the test is this Sigma-metric being estimated?

To continue with our arrow and target analogy, we know that our ability to hit the bull's-eye is not absolute, regardless of distance. Our bias and precision depend on the distance to the target (i.e., the analyte concentration). It is well recognized that analytical performance, as measured by bias and precision, varies with analyte concentration.

It is a common assumption that a method's performance can be summarized by a single statistic such as the Sigma-metric or a set of performance measures. But intuitively, we know that a method does not perform identically across its entire analytical range. At low, middle, or high analyte concentrations, performance may improve or degrade. Just as opera singers are categorized by their ability to sing in a certain vocal range (from bass to tenor for the men, from contralto to soprano for women), laboratory methods often perform better in certain parts of their analytical range. Manufacturers explicitly acknowledge this when they engineer high-sensitivity or ultra-high-sensitivity methods; usually this distinction indicates a method that has been optimized to provide better sensitivity (e.g., bias, precision) at the low end of its analytical range.

It is essential to measure method performance at levels where the use of the test is important. These levels are sometimes called critical levels, decision levels, medical decision levels (MDLs), cutoffs, cutoff thresholds, etc. While the terminology varies, the idea is the same: there are specific levels where test interpretation is critical for proper clinical interpretation. For our purposes, we will refer to these areas of test performance as MDLs. The most important point: if the method does not perform well at a MDL, clinical care may be compromised. Thus, we should measure performance and calculate Sigma-metrics at important MDLs.

Agreeing that we should measure performance where performance is important is hardly a revelation. Determining the MDLs for a given method, however, is more of a challenge.

There are few resources for MDLs that cover all assays, but here are some typical sources of information:

- Manufacturer controls and calibrators are typically targeted around MDLs
- External Quality Assurance (EQA), peer group, and proficiency testing samples will often send specimens or "events" that are also targeted to MDLs.
- Scientific literature often includes studies in which performanceat specific levels is measured, or where medical decisions and diagnoses are based on certain cutoffs. For example, tight glycemic control (TGC) is a protocol that has been discussed and disseminated through multiple publications and for which specific concentration limits are designated.
- Expert groups for certain diseases often set global guidelines for diagnosis and interpretation. For example, the American Diabetes Association (ADA) makes recommendations on how to interpret glucose and glycated hemoglobin (HbA1c) levels.
- Specific labs, hospitals or health systems may create clinical pathways or treatment and interpretation guidelines that are unique to their operations.

To an extent, the resources available for MDLs mirror the resources available for determining the quality requirement for a test. This should not come as a surprise, since the two issues are intertwined. The performance demanded of a test method is dependent on where in the analytical range you are measuring it. The benefit of this close correlation is that we can use the Stockholm Consensus Hierarchy[1] as a tool to rate which sources of information are preferred for determining MDLs. In the priority list below, the models mentioned first in the hierarchy are preferred.

While there are many resources that can be consulted, and laboratory testing is continually evolving, the information about how tests are interpreted and what cutoffs are used is constantly changing. Often, the authors of scientific literature engaged in active debate about which MDL is important, or where to draw an appropriate cutoff.

For example, the recommended screening cutoff for diabetes was once 7.0% HbA1c, but was recently reduced to 6.5%. Likewise, cholesterol treatment used to begin at 240 mg/dL, but clinical recommendations have recently lowered that to 200 mg/dL. Professional judgment is needed to review these suggestions and evaluate which MDL is truly the critical one. Finally, there is the challenge of the "art" of medicine: while national guidelines may set MDLs, individual physicians may and often do choose their own idiosyncratic ways of interpreting test results and making diagnosis and treatment decisions. The laboratory will have to weigh the different sources of information and find common ground between guidelines and practices, ultimately adapting the decision levels to the reality of how local clinicians are using test results.

One target may not be enough.

The experienced laboratory professional can call to mind several laboratory tests for which there are multiple MDLs. For example, with glucose, there are important decision levels for hypoglycemia and hyperglycemia. There are also cutoff levels for glucose that are used to diagnose diabetes as well as determine when glycemic control is poor. Can a single Sigma-metric calculation adequately reflect the performance of a glucose method across all of these multiple decision levels? Can the performance of a glucose method be summarized by any single statistic?

When multiple MDLs exist for a test method, the ideal practice is to measure performance at each of those levels. In practice, a method that performs well at one decision level often performs well at other decision levels (good precision is contagious). However, at lower levels, performance is often a challenge for even the best methods, so care must be taken to measure performance at the lower end of the analytical range when decisions are made at that MDL.

Knowing MDLs is just the start, however. It is the equivalent of knowing where the target is, without knowing the size of the target, or whether we have any chance of hitting it.

1	Evaluation of the effect of analytical performance on clinical outcomes in specific clinical settings		
2	Evaluation of the effect of analytical performance on clinical decisions in general	a	data based on components of biological variation
		b	data based on analysis of clinicians' opinions
3	Published professional recommendations	a	from national and international expert bodies
		b	from expert local groups or individuals
4	Performance goals set by	a	regulatory bodies
		b	organizers of External Quality Assessment (EQA) schemes
5	Goals based on the current state of the art	a	as demonstrated by data from EQA or Proficiency Testing (PT) schemes
		b	as found in current publications on methodology

Table 1

ADVICE FOR MANUFACTURERS	Measure performance at as many MDLs as possible, drawing on the best available knowledge on the use and interpretation of the test. When it is not practical to measure performance at a multitude of levels, select the most important levels. Also, when making multiple measurements, spread them across the analytical range, so that there are Sigma-metrics for low and high, or low, middle, and high MDLs.
ADVICE FOR LABORATORIES	Determine which MDLs are important for your laboratory and the clinical pathways in use at your healthcare system. It may be that only a few decision levels are relevant to your patient population. For those MDLs, calculate the Sigma-metric of the method.

Quality requirements: How big is the target?

How much error is allowable for a given method? Getting an answer for this question is harder than it seems. The difficulty of obtaining an answer is compounded by conflicting models and differing terminology. At various times, quality requirements have been expressed as medically allowable (maximum) inaccuracy, medically allowable (maximum) imprecision, total error, etc.

A recent review of this issue by Dr. George G. Klee of the Mayo Clinic concluded:

> *"There is no consensus currently about the preferred methods for establishing medically necessary analytic performance limits.The various methods give considerably different performance limits."*[2]

The Sigma-metric approach is predicated on a quality requirement that is expressed as a Total Error Allowable (TEa), but does not specify how this quality requirement should be determined or obtained. Thus, a quality requirement could be obtained from the U.S. CLIA* proficiency testing criteria, the Royal College of Pathologists of Australasia (RCPA) guidelines, the Ricos et al. database on Desirable Specifications for Total Error based on within-subject biologic variation, an ISO standard, a peer group specification, or even a locally determined specification. The laboratory must choose which quality requirement to use, knowing that the choice of an appropriate quality requirement is critical to the utility of the Sigma-metric.

Again, the Stockholm Consensus Hierarchy provides a framework to rate which sources of information are preferred for quality requirements. Here are some examples:

- The recent German RiliBÄK[3] specifications for interlaboratory comparisons represent level 4 in the hierarchy from Table 1, as do the CLIA guidelines for proficiency testing, as well as other EQA and PT groups, such as NEQAS**, WEQAS***, DGKL[†], and QMP-LS.[††]

- The recommendations of the ADA for treatment and diagnosis of diabetes represent level 3. The critical values for creatinine/eGFR have been specified by the National Kidney Disease Education Program (NKDEP) and the American Association for Clinical Chemistry (AACC) Laboratory Working Group.

- The Ricos et al Desirable Specifications derived from within-individual and individual-to-individual biologic variation represent level 2.

To reach the highest level of quality requirements, you would need to perform in-depth studies of the effect of changes in analytical performance upon patient populations. Using patient population studies, it is possible to model the impact of increasing bias upon the diagnosis and treatment of patients. In addition, it is possible to simulate the impact of short-term variation of test value distributions upon the patient population, again noting the impact on diagnosis and treatment decisions. An even more sophisticated approach is to examine actual test ordering behavior in a large population of patients and clinicians, to establish the relationship between test values and frequency of follow-up procedures.[4]

ADVICE FOR MANUFACTURERS	Determine the quality requirements at as many MDLs as possible, drawing on the best available knowledge on the use and interpretation of the test. When it is not practical to determine multiple quality requirements, try to select the quality required at the most important decision levels. Similar to the choice of decision levels, it may be helpful to determine the quality required at different parts of the analytical range (low and high, or low, middle, and high). At the same time, ensure that the method will meet important requirements imposed by regulatory bodies.
ADVICE FOR LABORATORIES	Determine the quality requirements for MDLs that are important to your laboratory and the clinical pathways in use at your healthcare system. It may be that only one use of the test is important, and only that quality requirement is needed.

These patient population and modeling studies are not a trivial endeavor. Probably only the most sophisticated health systems, large hospitals, or reference laboratories have the data and expertise to conduct such studies. For the majority of laboratories, it is common to rely upon quality requirements already available in published resources.

The main lesson of the Stockholm Consensus hierarchy is to emphasize evidence and actual clinical practice over survey-derived goals. The reality of test usage trumps national guidelines and expert advice. If an individual laboratory knows the clinical pathways that are used by their clinicians, those can be "reverse engineered" via a clinical QC design model into a quality requirement for the method. Such a goal will be more important to achieve than a requirement from a proficiency testing survey, a published scientific recommendation from a group of experts, or a goal set by a national standards group.

Nevertheless, there are times when "lesser" quality requirements will take precedence over more evidence-based ones. For example, if examining clinical pathways in the local hospital leads to a quality requirement that is larger than the quality requirement imposed by an EQA or proficiency testing program, the smaller goal will necessarily prevail. Regulatory compliance is often compulsory; when requirements are enforced by the government or laboratory accreditation authorities, they must be met, even if the requirements are not the most clinically appropriate. Usually, however, the requirements set by more evidence-based approaches are smaller, not larger, than those set by regulatory bodies. From a global perspective, this means that the country with the most demanding and stringent quality requirements may unwittingly establish the benchmark for compliance and performance.

Remember, these quality requirements should be determined at MDLs, where the performance of the test method is relevant to the clinical pathways.

Having found where the target is, as well as the size of the target, our task is still not complete. We need to get the best information available to determine whether or not we can hit that target with our method performance.

Getting the best estimate for imprecision

Imprecision is one of the few factors that is entirely "local" and within the power of the laboratory to measure. That is not to say that the laboratory has complete control over method precision – the performance of the instrument depends heavily on the engineering. But while some of the other factors in the Sigma-metric approach are set by external factors (for example, the local use of the test by clinicians may define the quality requirement, or conversely, the quality requirement may be set by a regulatory body that is completely outside the health system), precision can be measured and monitored just by using the resources found within the laboratory.

The core idea behind the measurement of imprecision is to get an estimate that reflects as closely as possible the real performance of the method in daily operation. This means taking into account the effect of different operators, different control lots, different reagent lots, and even the difference in operation between weekdays and weekends.

While manufacturers frequently provide the skilled staff to verify method performance of a new instrument during its installation in the laboratory, it is preferable to have the actual laboratory staff perform the precision studies. In this way, the estimate of imprecision reflects the performance of the true operators, not the performance of transient field technicians.

Page 7 lists the ranking of preferred estimates of imprecision. Following the pattern established earlier with the Stockholm Consensus hierarchy, the preferred sources of data are listed first.

1	Routine, historical imprecision, sometimes called cumulative coefficient of variation (%CV), measured over a long term. This may represent the summarized data of several months of routine control data. The CLSI C24[5] guideline recommends three to six months of routine data for a calculation of %CV.
2	Total imprecision, sometimes called intermediate precision. Typically, a total precision study is conducted, which consists of two runs per day of a sample, run for 20 days. The CLSI EP5[6] guideline specifies how to conduct a total precision study and calculate the Total Imprecision.
3	Within-day or between-run imprecision. Two runs within a single day, each run consisting of 10-20 replicates of a sample. This is an improvement on the within-run estimate, since it expands the coverage to more than one run. However, this estimate may be too optimistic.
4	Within-run imprecision, sometimes called repeatability. This is the easiest type of study and is conducted frequently, particularly in method validation studies. Typically performed within a single run, with at least 20 replicates of a sample. This imprecision estimate reflects a narrow window of actual performance. Often it provides an excessively optimistic estimate of imprecision, as it does not include all the sources of variation that the method will routinely experience.

Table 2

ADVICE FOR MANUFACTURERS	Conduct a large scale precision study or, at a minimum, conduct a Total Precision study using CLSI guidelines. The more data, the better. Conduct these precision studies at relevant MDLs. When it is not practical to measure imprecision at a multitude of levels, try to select the most important MDLs. Also, when making multiple measurements of imprecision, spread them across the analytical range, so that there are Sigma-metrics for low and high, or low, middle, and high.
ADVICE FOR LABORATORIES	If possible, use your records to look at several months of routine control data and calculate a cumulative CV. At a minimum, conduct a Total Precision study following the CLSI EP5 guideline. The more data, the better. Again, remember to conduct these precision studies at MDLs that are important for your laboratory.

Getting the best estimate for bias.

"Analytic bias caused by assay differences and reagent variations can cause major problems for clinicians trying to interpret the tests results."[7]

While determining precision is relatively straightforward, measuring bias can sometimes take the form of an existential debate. How do we measure bias? Bias is a relative term – we measure it against something else – so we must decide: What is the standard for comparison?

There are many possible methods to calculate bias, which will be discussed first. But in addition to the technique of calculating the bias, there is the additional complication of determining the standard against which the new method is judged. As discussed in the previous sections, we can rank the preferred source of biases in a hierarchy. In the list below, methods of calculating bias mentioned first are preferred:

- Bias from reference material or reference method
- Bias from the mean of a peer group
- Bias from the all-method mean of a proficiency testing (PT) or EQA survey
- Bias from a comparative method

The preferred technique of measurement and estimation of bias

The CLSI EP9[8] guideline details the Comparison of Methods study, which typically uses a minimum of 40-60 patient specimens spaced across the reportable range, measured on both the new method and the comparison method, over a recommended time period of at least 5 days. Comparison and difference plots may be used to assess the bias. Regression statistics, whether they are linear, Deming, or Passing-Bablok, can be used to generate the regression equation, with its slope and y-intercept. The regression equation can then be used to calculate the bias estimate specifically at the MDL where quality requirements Total Error Allowable (TEa) have been specified and precision has been measured.

1. Calculating bias from a reference material, reference method, or standard

For some analytes, there is a gold standard (or reference) method or material. There is, in other words, a "true" value that should be achieved by all methods. To get to this true value, and relate our laboratory method to it, we must enter the world of Metrology.

"Metrology has been very good about identifying reference methods and reference materials and putting together a formal traceability chain so that you can tie your kit calibrator in your clinical lab back to a reference material and a reference method that are internationally recognized… The whole idea is that you can then come close to scientific truth rather than a test result that is a relative truth." [9]

Bias is very easy to estimate: the result of the test method is subtracted from the "true" value determined by a comparison or reference method. When we calculate bias using a reference method and/or reference material, we're figuring out a "true" bias, a bias that is more scientifically true than just relatively true. When we determine bias from a reference method, we are calculating how far away our test result is from the true answer. When we determine bias from a comparison method, we only calculate how far away we are from a different answer.

For laboratories seeking reference methods, the Joint Committee for Traceability in Laboratory Medicine (JCTLM) database[10] has a listing of reference methods which can provide target values and methods that represent the scientific true values.

2. Bias from a peer group

The next possible way to measure bias is through a peer group. This is very similar to participating in PT or an EQA program, except all the participants in the testing event use similar instruments and methods. Typically, a peer group is a group of labs that use the same instruments and the same controls and/or reagents. So the answers that each laboratory obtains should be much closer to each other. Again, while there should be smaller differences between participants, the peer group mean is not a "true" mean such as obtained with a reference method and/or reference material. Peer group means are, in effect, all-method means for a single method. There is more confidence that a bias exists, because if a laboratory's value varies from all of the peer values, there must be an issue (we can't blame the difference on different methods or materials anymore). If the peer group is also using a reference material or including a reference method measurement with the results, the value of the report is improved. Still, in the absence of additional information, peer group reports are useful, but they cannot tell us if we have a "true" bias.

3. Bias calculated from PT or EQA

One of the routine ways to determine bias is to compare the results of our laboratory against those of other laboratories through PT or EQA. Typically, a sample is sent out to all laboratories in the program, all laboratories run the sample and report the result, then the program tabulates the results and issues a report back to the labs. Each report typically states the difference (or bias) between the individual laboratory's result and that of the PT/EQA group method mean. Given that information, each individual laboratory is supposed to decide if the bias is significant and warrants a correction, adjustment, or calibration on their part.

For some analytes, reference methods and/or reference materials are used, so they include a definitive "true" value for the event or sample. This means that all labs should get a specific result, and every increment away from that result is considered "true" bias. If we determine bias using the method mean from a reference method or reference material, we are measuring the difference between our result and the "true" result.

For many other analytes, no reference methods exist – or, even though a reference method may be available, the PT/EQA group might not test the sample using it and establish a target value – so there is no definitive value reported. Instead, the report only states the "all-method" mean. There are other terms for this mean; sometimes it is simply called the "group mean." Essentially, this mean is the average of all the different laboratory results (albeit trimmed in some way). In other words, the mean is close to the answer that all the laboratories reported. This doesn't mean that the answer is the "true" answer, because all the laboratory methods could be biased in the same direction (revisit the concept of "precise but not accurate"). Here, if we determine bias using an all-method mean, we are measuring the difference between our laboratory result and the results that most of the other laboratories obtained.

4. Bias from a comparative method

Part of the method validation process includes a method comparison study, typically done between the new method that has just been purchased and the old method which is being replaced. Note the difference between “comparative” and “reference” method. The comparative method is only a relative comparison – there is no claim to scientific truth here. It could be that the old method was more scientifically true while the new method is less scientifically true, so a new relative bias exists in the wrong direction.

However, even if not “true,” a bias determined by a comparison study can still be quite real – because patient care and test results that span the switch-over to the new method will be shifted up or down, even when there is only a relative difference between the new and old methods. A patient receiving care before and after the switch could see a rise or fall in their test values, resulting, in the worst case, in misdiagnosis and treatment. So this is a “real” bias – even if it isn’t “true” bias.

There is a special case of this type of bias calculation for diagnostic manufacturers. Since manufacturers often produce instruments in various sizes – from small point-of-care devices to small office laboratory instruments, to medium-sized hospital instruments to core laboratory, high-volume, automated instruments – there is a product “family” within which customers are reasonably expected to operate. The common practice of modern health systems is to use instruments from the same “family” across their whole system, so it is logical to assume that the small-volume instrument will be used in conjunction with the core laboratory instrument of the same manufacturer. There could be a real bias problem if customers using instruments of the same product line had significant differences between the test results. Thus, manufacturers find it important to measure biases between their instruments and ensure that these differences are not analytically or clinically significant.

A note of caution: absolute bias, relative bias, and relevant bias

The determination of bias presents many challenges for laboratories and manufacturers alike. It is not difficult to agree that laboratories should compare performance of their methods against standard reference methods, or reference materials, or something that is as high as possible on the traceability chain. This ideal choice, unfortunately, may not always be practical or even possible. There are numerous methods on the market for which reference methods simply do not exist, where assays are not standardized or even harmonized, or where the cost and/or availability of good reference methods render them out of reach of the typical laboratory.

In cases where the ideal is not possible, manufacturers and laboratories should consult the hierarchies in Table 1 and attempt to use the next best source for that component.

ADVICE FOR MANUFACTURERS	The relevant bias is twofold: (1) what is the bias of the new method from a standard reference method or material? and (2) what is the bias of the new method from methods that are within the current or past product family? Frequently, laboratories and health systems select complete product lines for installation, and use these products across their entire system, so manufacturers should design methods that aim for the smallest possible bias to a standard reference method, but also balance the differences between their different instrument methods. Again, it is important to identify the MDLs and determine the bias at each of these levels.
ADVICE FOR LABORATORIES	Here, the strategy is slightly different. During instrument selection, it is paramount to select methods that have minimal bias when compared to standard reference methods, particularly for the largest instruments found in the core laboratory. Once a method that is closest to the “true” values is selected and installed, the focus shifts. All the other subsidiary methods of the health system should then adjust their calibrations to minimize the bias against the core laboratory methods. In effect, the core laboratory instrument becomes the “local” standard reference method. Smaller instruments and methods in secondary and tertiary laboratories should measure and minimize their bias against the core laboratory instrument.

There is also the question of relevant bias. As discussed earlier, comparing the new laboratory method to a standard reference method is ideal, but if the new laboratory method is replacing an old method, the difference between those two field methods is probably more relevant to the local clinical care. For patients whose records and care span from the time of the old method into the installation and use of the new method, their records will explicitly reflect the bias between the methods. Whether or not that bias is "true" is not as important. Sometimes the relative bias is the more meaningful bias.

It follows logically that hospitals and large health systems will seek similar instrumentation across their organization, based on the premise that diagnostic manufacturers create product lines with closely aligned methods. A well-engineered instrument family will have smaller biases between instruments. When this closeness of agreement among methods and instruments can be confirmed, the selection of an entire product line is justified and operational gains can be achieved in the health system. A heterogeneous mix of instruments, in which each device and method has different engineering and possibly even different measurement principles, could in theory produce significant biases between the instruments. Examples of bias between instruments of different manufacturers are easy to find – just look at any available proficiency testing survey. Within a health system, method validation studies can determine the extent of these biases using the comparison of methods experiment (or possibly split-sample studies). Sigma-metrics can be used to highlight the impact of these biases and summarize the performance of methods within the health system. The Sigma-metric approach allows comparison of methods with objective single quantitative measures of analytical quality.

Warning: your mileage – your metric – may vary

After noting all the different challenges we're faced with when gathering each component of the Sigma-metric equation, it may seem like an impossible challenge to obtain a consistent set of Sigma-metrics. At every stage, there are choices to make. Ultimately, many of the calculations are specific to the local context of the laboratory. The choice of quality requirement may also be strongly influenced by the regulatory and accreditation context. Bias calculations in particular will depend upon what other methods exist in the health system.

The challenge is harder still when it comes to examining Sigma-metrics that are published in the literature, or claims of Sigma-metrics made by manufacturers. Particularly because the choice of quality requirement is not standardized, simply knowing the Sigma-metrics from a study is not sufficient. It is imperative to know which goals were used to calculate those metrics, and how the performance of the method was measured and estimated. Knowing the final number is not enough; the laboratory must also understand the "ingredients" of that metric.

It will be tempting to construct optimistic estimates of imprecision and inaccuracy, to choose the easiest target to hit, to measure performance at analyte concentrations where the numbers are impressive and/or available rather than MDLs where the clinical interpretation is critical. Laboratories may be tempted to "goal-shop" to find the largest quality requirement, which then provides them with the easiest target to hit. Numerically, this is an understandable choice. If we pick the largest numerical target, our Sigma-metric will of course be the highest number. Clinically, however, this may be detrimental to patient care. Choosing a large target may generate impressive Sigma-metrics, but if the actual clinical use of the test is dependent upon hitting a much smaller target (i.e., achieving a more demanding level of performance), the real Sigma-metric performance of our test is different than our delusional metric.

Picking a target larger than the clinically optimal target actually required for appropriate care cheats the laboratory and, ultimately, impairs the clinician and patient.

Building a false Sigma-metric benefits no one. For manufacturers, a misleading Sigma-metric will ultimately result in unhappy customers, who experience less quality than was promised and is subsequently expected in routine operation. For laboratories, a falsely elevated Sigma-metric may result in reduced QC effort (fewer rules, wider limits, possibly even reduced QC frequency) and lull the staff into complacency. Meanwhile, the necessary QC is not being performed and medically important errors may go undetected.

The advantage of a statistic such as the Sigma-metric is that in one number, it neatly summarizes a characteristic of multiple key analytical performance characteristics. In other words, a single number can say a lot more when it is a clinically meaningful and useful statistic, rather than a number randomly selected or generated. Just as the mean or average of a set of data (which is reasonably normally distributed) tells us a lot more than just one of the numbers from that data set, a Sigma-metric can tell us a lot more than just the results from one of the method validation studies.

However, the strength of a statistic can also be its weakness. Because it is a single number, it has the disadvantage of being hard to understand without the proper background or explanation or context. This is why statistics are so easily misused and distorted, because once we extract the statistic from the data and environment in which it was created, it can be hard to confirm that the number has any useful meaning. For those individuals with less statistical knowledge, statistics can often be as confusing as they can be illuminating.

So statistics answer questions, but they raise them as well. If a statistic seems to pass judgment on the data, then the next question becomes: was the data correctly generated and analyzed? Were the statistics calculated correctly? Does this statistical finding apply only in theory but not in practice? Is this finding relevant to our situation?

With Sigma-metrics, whether we read about them in the literature or hear about them from another laboratory, we are faced with a conundrum. If we are told, or if we read somewhere, that the Sigma-metric of a cholesterol test is 6, what does that really tell us? It may indicate excellent performance and world class quality, but the wise laboratory will want to learn more, including how the metric was calculated (i.e., at what MDL), with what quality requirement (TEa), with what estimates of imprecision and bias, and against what comparison or reference standard the bias was calculated.

The quality of a quality metric matters

Given the effort and resources required to determine the appropriate quality requirements at appropriate MDLs and obtain the most realistic estimates for imprecision and bias, one may be tempted to question whether this is worthwhile.

We want to put the same degree of effort into our quality metrics as we put into the laboratory processes themselves. The quality of our quality metrics must be as high as the quality of our laboratory. Using a Sigma-metric of high reliability, we can have increased confidence in the performance of our laboratory processes.

The reliability of a quality metric lays the bedrock for the quality of the test. The quality of the test then provides solid ground for the diagnosis by the clinician and treatment of the patient. By following best practices for determining Sigma-metrics, we give the clinician a sturdy foundation, so they can hit the bull's-eye for patient care.

*CLIA: Clinical Laboratory Improvement Amendments 1988, USA

**NEQAS: National External Quality Assurance System, UK

***WEQAS: Welsh External Quality Assurance System, Wales

†DGKL: Deutsche Gesellschaft Für Klinische Labor, Germany

††QMP-LS: Quality Management Program - Laboratory Service, Ontario, Canada

1. Consensus agreement, Strategies to Set Global Quality Specifications in Laboratory Medicine, Stockholm, April 24-26, 1999. http://www.westgard.com/stockholm.html
2. Klee GG. Establishment of Outcome-Related Analytic Performance Goals. *Clinical Chemistry*, 2010; 56:(5):714-722.
3. RiliBÄK. Unofficial English Translation. http://www.westgard.com/rilibak.htm.
4. Karon BS, Boyd JC, Klee GG. Glucose Meter Specification Criteria for Tight Glycemic Control Estimated by Simulation Modeling. *Clinical Chemistry,* 2010;56:1091-1097.
5. CLSI Document C24-A3. Statistical Quality Control for Quantitative Measurement Procedures. CLSI, 940 West Valley Road, Wayne, Pennsylvania, 2002.
6. CLSI EP5-A. Evaluation of Precision Performance of Quantitative Measurement Methods. CLSI, 940 West Valley Road, Wayne, Pennsylvania, 2004.
7. Klee GG. Clinical interpretation of reference intervals and reference limits. A plea for assay harmonization. *Clinical Chemistry and Laboratory Medicine*, 2004; 42(7):752-757.
8. CLSI EP9-A. Method Comparison and Bias Estimation Using Patient Samples. CLSI, 940 West Valley Road, Wayne, Pennsylvania, 2002.
9. David Armbruster, quoted in The Pursuit of Traceability, written by Bill Malone. *Clinical Laboratory News*, October 2009, cover story.
10. Joint Committe for Traceability in Laboratory Medicine (JCTLM) database of higher-order reference materials, measurement methods/procedures and services. http://www.bipm.org/jctlm/ Accessed July 16, 2010.

Sten Westgard received financial compensation from Abbott Diagnostics for his work on this paper.

ABBOTTDIAGNOSTICS.COM

Abbott