The Quality of Laboratory Testing Today

An Assessment of σ Metrics for Analytic Quality Using Performance Data From Proficiency Testing Surveys and the CLIA Criteria for Acceptable Performance

James O. Westgard, PhD,[1,2] *and Sten A. Westgard, MS*[2]

Key Words: Quality assessment; Clinical Laboratory Improvement Amendments; CLIA; Proficiency testing; Quality requirements; σ metrics

DOI: 10.1309/V50H4FRVVWX12C79

Abstract

To assess the analytic quality of laboratory testing in the United States, we obtained proficiency testing survey results from several national programs that comply with Clinical Laboratory Improvement Amendments (CLIA) regulations. We studied regulated tests (cholesterol, glucose, calcium, fibrinogen, and prothrombin time) and nonregulated tests (international normalized ratio [INR], glycohemoglobin, and prostate-specific antigen [PSA]). Quality was assessed on the σ scale with a benchmark for minimum process performance of 3 σ and a goal for world-class quality of 6 σ. Based on the CLIA criteria for acceptable performance in proficiency testing (allowable total errors [TE_a]), the national quality of cholesterol testing (TE_a = 10%) estimated σ values as 2.9 to 3.0; glucose (TE_a = 10%), 2.9 to 3.3; calcium (TE_a = 1.0 mg/dL), 2.8 to 3.0; prothrombin time (TE_a = 15%), 1.8; INR (TE_a = 20%), 2.4 to 3.5; fibrinogen (TE_a = 20%), 1.8 to 3.2; glycohemoglobin (TE_a = 10%), 1.9 to 2.6; and PSA (TE_a = 10%), 1.2 to 1.8. The analytic quality of laboratory tests requires improvement in measurement performance and more intensive quality control monitoring than the CLIA minimum of 2 levels per day.

What is the quality of laboratory tests today? Studies of laboratory errors have documented that a higher percentage of errors occur in the preanalytic and postanalytic processes than in analytic processes.[1-3] The figures often quoted are 45% for errors in preanalytic processes, 10% for analytic errors, and 45% for postanalytic errors (actual estimates, 45.5%, 7.3%, and 47.2%, respectively) based on a study done in 1988[1] before the implementation of the current Clinical Laboratory Improvement Amendments (CLIA) regulations. As a consequence of this expected distribution of errors, laboratories are urged to focus their attention on preanalytic and postanalytic processes to improve patient safety.[4]

The final CLIA rule reflects this emphasis on increased quality assessment for preanalytic and postanalytic processes[5] and proposes a reduction in quality control (QC) for analytic processes. This proposal for reducing QC is not found in the regulations but in the State Operations Manual (SOM),[6] which provides interpretive guidelines for implementing the regulations. According to the SOM, laboratories may be able to reduce QC from 2 levels per day to 2 levels per week or even 2 levels per month for measurement procedures and instruments with built-in controls. These so-called equivalent QC (EQC) procedures would be particularly attractive for point-of-care testing applications in which operators often have little laboratory experience and minimum analytic skills.

Is the analytic quality of laboratory tests really so good that only weekly or monthly QC is needed? There are few studies that document current analytic performance relative to the quality required for medical usefulness. The Centers for Medicare & Medicaid Services' (CMS's) own data from laboratory inspections show that as many as 5% to 10% of laboratories are deficient in QC practices,[7] which should raise concerns about the

analytic quality achieved. Under CLIA, laboratories also must participate in proficiency testing (PT) surveys, which provide another source of data about the quality of laboratory testing. These data make it possible to provide a more quantitative assessment of the "state of the art" of laboratory testing.

For assessing quality in other industries, Six Sigma Quality Management is gaining momentum as the best approach for providing objective estimates and metrics.[8] Six Sigma requires that *tolerance limits* be defined for good quality to objectively identify poor quality or defective products (or erroneous test results). Two methods exist, one that inspects process outcome and counts the defects, calculates a defect rate per million, and uses a statistical table to convert defect rate per million to a σ metric. The second makes use of estimates of process variation to predict process performance by calculating a σ metric from the defined tolerance limits and the variation observed for the process. The first method is applicable to preanalytic and postanalytic processes, whereas the second method is particularly suitable for analytic processes in which the precision and accuracy can be determined by experimental procedures. Nevalainen et al[9] demonstrated the application of Six Sigma concepts for characterizing the quality of preanalytic and postanalytic processes on the σ scale. Applications to analytic processes have been described by Westgard.[10]

When assessing quality on the σ scale, the higher the σ metric, the better the quality. According to Nevalainen et al,[9] "average products, regardless of their complexity, have a quality performance value of about 4 σ. The best, or 'world class quality,' products have a level of performance of 6 σ." This corresponds to a process capability index of 2.0, which also has been the goal for industrial production processes.[11] Industry recommends a minimum acceptable process capability of 1.0,[12] which would correspond to a 3 σ process. Thus, common goals across industries are to strive for 6 σ quality and accept a minimum of 3 σ quality. As might be expected, the size of analytic errors that need to be detected by QC will depend on the process capability[13]; therefore, the σ metric also is useful for assessing the adequacy of QC procedures and practices.[10] Thus, with the aid of Six Sigma principles and metrics, it is possible to assess the quality of laboratory testing processes and the QC that is needed to ensure that the desired quality is achieved.

Materials and Methods

Tolerance Limits or Quality Requirements

For laboratory tests, external quality assessment and PT programs provide readily available sources of tolerance limits, most often stated as allowable error limits or allowable total errors. An allowable total error (TE_a) encompasses the imprecision and bias of a single test measurement; thus, it fits the desired form of a tolerance limit. Other types of quality requirements, such as biologic goals for imprecision and bias, may be converted into biologically allowable total errors.[14] Clinical outcome criteria, such as a decision interval for test interpretation, require that preanalytic variables, as well as biologic variation, be taken into account[15,16]; therefore, applications of clinical requirements are more complicated.

Selected Laboratory Tests

This assessment focuses on a few tests—cholesterol, glucose, calcium, glycohemoglobin, prothrombin time, international normalized ratio (INR), fibrinogen, and prostate-specific antigen (PSA). Some are regulated analytes for which CLIA specifies criteria for acceptable performance (TE_a) in required PT events. For example, cholesterol should be correct within 10%; glucose within 6 mg/dL or 10%, whichever is greater; calcium within 1.0 mg/dL; prothrombin time within 15%; and fibrinogen within 20%. For INR, the College of American Pathologists (CAP) sets a quality requirement of 20%. For glycohemoglobin and PSA, which are nonregulated tests that are widely used and have important medical applications, a "sensitivity assessment" can demonstrate the quality available from routine testing methods.

PT Programs

CMS identifies 14 approved PT providers. We selected 5 on the basis of the types of laboratories they serve and the availability of PT results: (1) American Academy of Family Physicians, which serves family physician office laboratories; (2) Medical Laboratory Evaluation, part of an alliance with the American College of Physicians and serves physician office and group practice laboratories; (3) American Association of Bioanalysts, which identifies itself with community clinical laboratories that serve clinic laboratories and small hospital laboratories; (4) American Proficiency Institute, one of the largest providers, with more than 12,000 clients many of which are small and medium-sized hospital laboratories; and (5) CAP, the major accrediting service for large hospital laboratories.

σ Quality Metrics

Three estimates of quality were calculated from the PT data. These estimates differ in how they account for random and systematic errors, or imprecision and bias, of the measurement procedures.

National test quality (NTQ) is calculated from the CLIA total allowable total error (TE_a) divided by the group SD or coefficient of variation (CV), ie, NTQ σ = TE_a/CV_{group}. This estimate of quality includes the random and systematic errors as part of each test result from every

DOI: 10.1309/V50H4FRVVWX12C79

participating laboratory. It is particularly relevant for tests that are interpreted against national treatment guidelines with national reference limits or national cutoff points. To pool the estimates from different survey programs, the average NTQ is weighted for the proportion of laboratories represented by the different survey programs.

National method quality (NMQ) is calculated from the CLIA requirement and the bias and CV determined for each method subgroup, ie, NMQ $\sigma = (TE_a - bias_{methsubgroup})/CV_{methsubgroup}$. This estimate also is applicable for national guidelines for reference limits and cutoffs, but this calculation first characterizes the σ quality of each method subgroup and then provides a weighted average for the survey to account for the proportion of laboratories in the respective subgroups. Finally, the average NMQ for all laboratories is estimated by weighting each survey average on the basis of the proportion of laboratories represented by the different survey programs.

Local method quality (LMQ) is calculated from the CLIA requirement and the CV determined for each method subgroup, without accounting for method bias, ie, LMQ $\sigma = TE_a/CV_{methsubgroup}$. For this estimate of quality to be relevant, it is necessary that laboratory tests be interpreted against locally determined reference intervals or medical cutoffs to compensate for any method bias present. The average for each survey is weighted for the proportion of laboratories in each method subgroup, and then the overall average for multiple surveys is weighted for the number of laboratories represented in each survey program.

σ Metrics QC Assessment Tool

The quantitative relationship between the quality of a measurement procedure, as characterized by its σ metric, and the appropriate QC procedure(s) can be shown by imposing a σ scale on a critical-error graph, a QC planning tool that has been used for many years. The critical error graph makes use of a process capability index[11] called the critical systematic error (ΔSE_{crit}), which is related to σ as follows:

$$\Delta SE_{crit} = [(TE_a - bias)/s] - 1.65 = \sigma - 1.65$$

Figure 1 shows a critical-error graph with a σ scale imposed. The probability of rejecting an analytic run is shown on the y-axis vs the size of systematic error on the x-axis (bottom scale) or the σ metric (top scale). The power curves, from top to bottom, represent the QC procedures shown in the key, from top to bottom. Given the objective of achieving a probability of error detection of 0.90 (or 90% chance), a testing process having 6 σ quality can be easily controlled with 3 SD control limits and 2 control measurements per run. A 5 σ process would be better controlled using 2.5 SD control limits. At 4 σ, it is necessary to increase the number of control

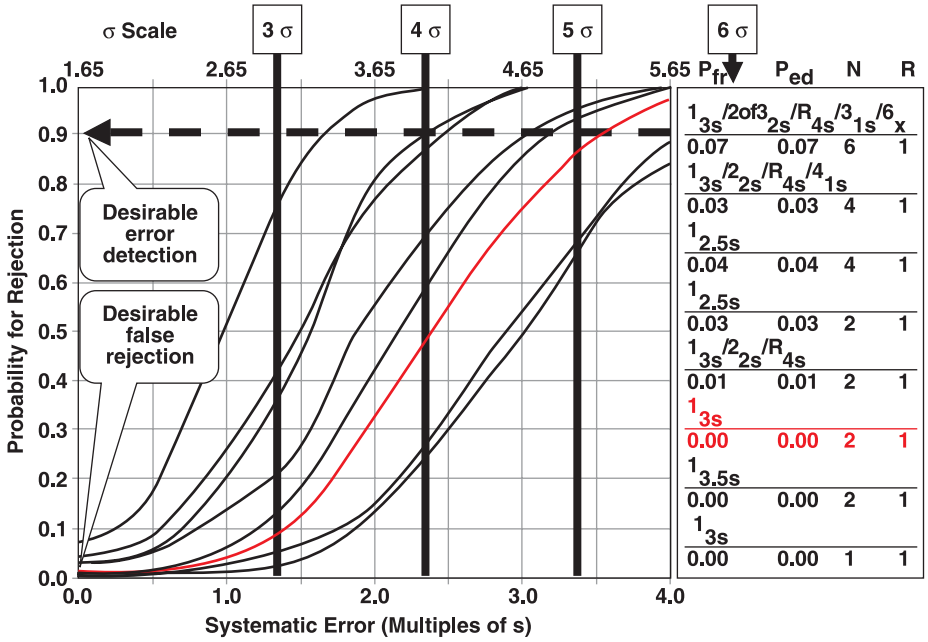

Figure 1 σ quality control assessment tool. The probability for rejection is shown on the y-axis vs the σ scale on the x-axis (top) and critical systematic error (bottom). The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. s, standard deviation.

DOI: 10.1309/V50H4FRVVWX12C79

measurements to 4 per run, and it becomes advantageous to use a multirule QC procedure to maximize error detection. At 3 σ, it is very difficult to ensure that analytic quality is satisfactory, even with 6 control measurements and multirule criteria.

Results

Table 1 shows the estimates of quality on the σ scale for cholesterol, calcium, glucose, and glycohemoglobin. These estimates represent 9,258 laboratories participating in PT for cholesterol, 9,786 for calcium, 10,722 for glucose, and 5,066 for glycohemoglobin. Approximately half of the laboratories for each of these tests are from the CAP survey program, thus the average σ figures for each test, shown in bold, are influenced most heavily by the CAP results. As expected, the estimates for NTQ are the lowest, being 2.88 σ for cholesterol (TE_a = 10.0%), 2.84 for calcium (TE_a = 1.0 mg/dL), 2.95 for glucose (TE_a = 10.0%), and 1.93 for glycohemoglobin (TE_a = 10.0%). The σ values for the NMQ were a little higher (3.02 for cholesterol, 3.00 for calcium, 3.34 for glucose) except for glycohemoglobin, which was the same. The σ values for LMQ were the highest and most optimistic: 3.67 for cholesterol, 3.86 for calcium, 4.00 for glucose, and 2.57 for glycohemoglobin.

Figure 2, **Figure 3**, **Figure 4**, and **Figure 5** show these estimates of quality in relation to the performance characteristics of different QC procedures. For cholesterol (Figure 2), even with the most optimistic estimate of quality (LMQ), laboratories need to use a multirule procedure with at least 4 control measurements per run. For more realistic estimates of quality (NTQ and NMQ), even more QC is needed, 6 control measurements per run with multirule criteria. Figure 3 reveals similar findings for calcium. Figure 4 shows somewhat better quality for glucose but still reveals the need for at least 4 control measurements per run. Figure 5 shows that glycohemoglobin testing has a more serious problem with quality and that current methods are not controllable to the same quality as required for glucose.

Table 2 provides a sensitivity assessment for glycohemoglobin and PSA to demonstrate the σ metrics achievable with different quality requirements. For uniform national guidelines to be applicable for interpretation of glycohemoglobin tests, today's methods are reliable only for an allowable total error of 2.0 to 2.5 %hemoglobin (Hb), or approximately 25% at a critical decision concentration of 7.0 to 8.0 %Hb. For

Table 1
Estimates of Analytic Quality for Cholesterol, Calcium, Glucose, and Glycohemoglobin as Determined From National PT Surveys*

PT Program	No. of Laboratories	Group Mean	NTQ (σ)	NMQ (σ)	LMQ (σ)
Cholesterol with TE_a = 10.0%					
AAFP	296	201.0	2.01	2.01	2.54
MLE	577	224.4	2.27	2.38	2.99
AAB	1,498	223.0	2.37	2.68	3.51
API	2,647	221.3	2.28	2.37	3.19
CAP	4,240	198.7	3.57	3.71	4.19
Summary	**9,258**	**210.8**	**2.88**	**3.02**	**3.67**
Calcium with TE_a = 1.0 mg/dL					
AAFP	164	10.2	2.50	2.35	2.71
MLE	528	10.5	2.44	2.69	3.50
AAB	1,444	11.1	2.78	2.95	3.37
API	2,695	11.1	2.63	2.98	3.45
CAP	4,955	10.4	3.03	3.07	4.30
Summary	**9,786**	**10.7**	**2.84**	**3.00**	**3.86**
Glucose with TE_a = 10.0%					
AAFP	245	134.0	1.91	2.64	3.16
MLE	628	106.1	1.75	2.13	2.99
AAB	1,665	106.4	2.22	2.60	3.20
API	3,038	106.6	2.42	2.70	3.24
CAP	5,146	149.6	3.70	4.14	4.88
Summary	**10,722**	**120.5**	**2.95**	**3.34**	**4.00**
Glycohemoglobin with TE_a = 10.0%					
AAFP	209	9.30	1.82	2.12	2.76
MLE	342	9.03	1.31	1.15	2.33
AAB	885	8.11	1.53	1.82	2.50
API	1,650	9.27	1.69	1.69	2.35
CAP	1,980	9.30	2.43	2.29	2.82
Summary	**5,066**	**9.06**	**1.93**	**1.93**	**2.57**

AAB, American Association of Bioanalysts; AAFP, American Academy of Family Physicians; API, American Proficiency Institute; CAP, College of American Pathologists; LMQ, local method quality; MLE, Medical Laboratory Evaluation; NMQ, national method quality; NTQ, national test quality; PT, proficiency testing; TE_a, allowable total errors.
* Presented as σ metrics. The group means for cholesterol, calcium, and glucose are given in conventional units (mg/dL); for glycohemoglobin, as the percentage of hemoglobin. The following are conversion factors for Système International units: cholesterol, multiply by 0.02586 (mmol/L); calcium, multiply by 0.25 (mmol/L); glucose, multiply by 0.05551 (mmol/L). Summary figures, in bold, are weighted averages that account for the relative number of laboratories in the respective PT groups.

DOI: 10.1309/V50H4FRVVWX12C79

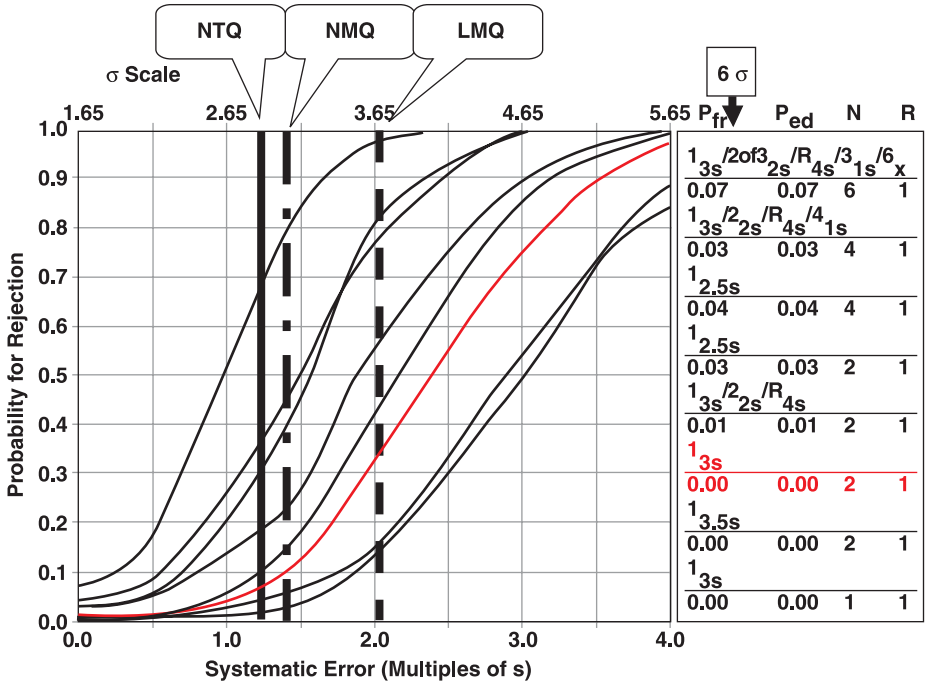

Figure 2 Cholesterol quality in US laboratories for a TE_a of 10%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

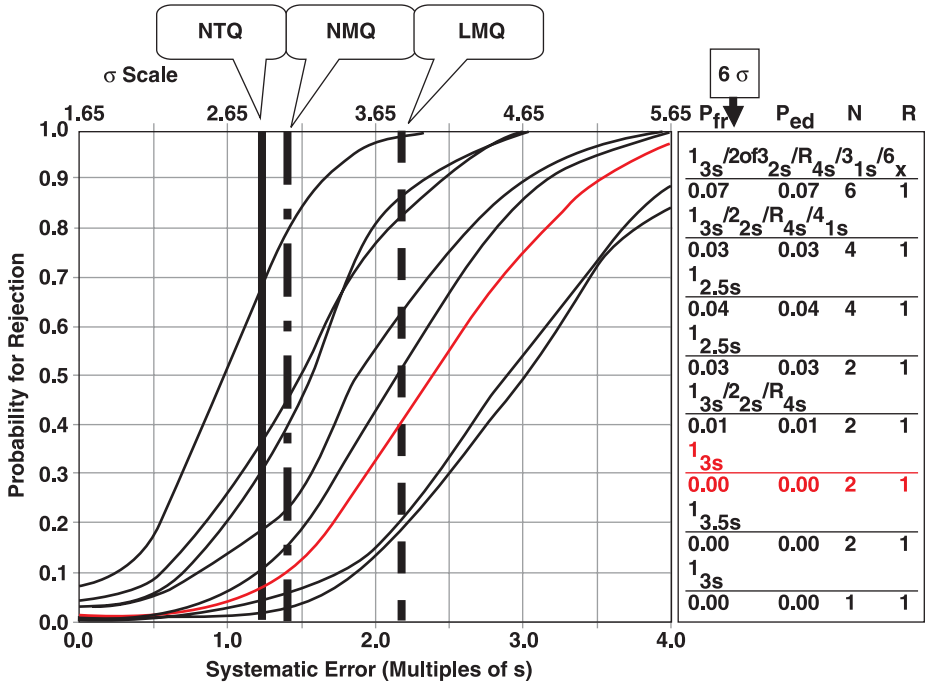

Figure 3 Calcium quality in US laboratories for TE_a = 1.0 mg/dL. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

DOI: 10.1309/V50H4FRVVWX12C79

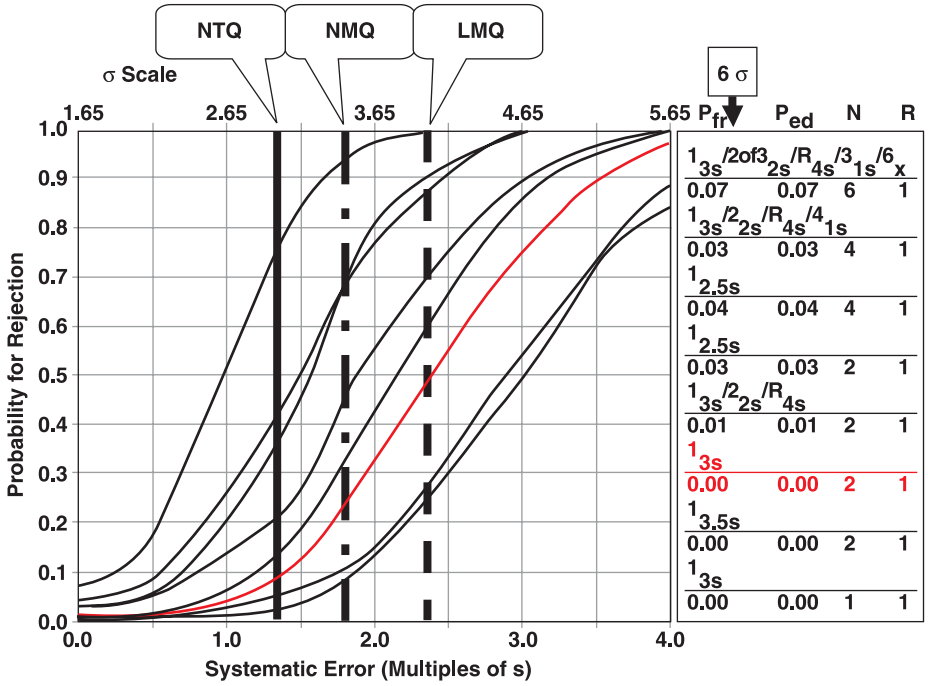

Figure 4 Glucose quality in US laboratories for TE_a = 10%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

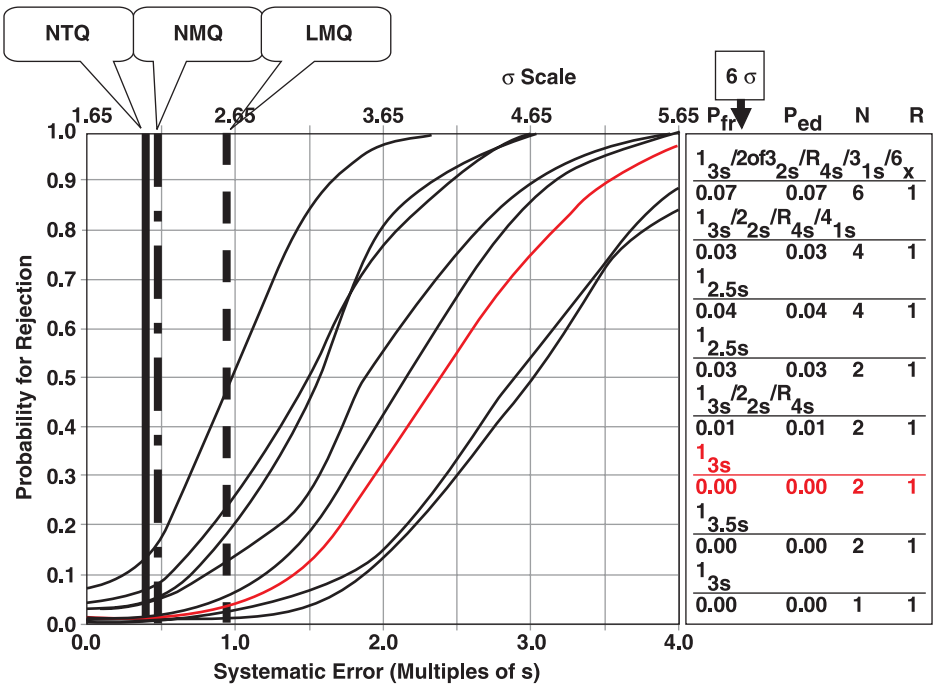

Figure 5 Glycohemoglobin quality in US laboratories for TE_a = 10%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

DOI: 10.1309/V50H4FRVVWX12C79

Table 2
Estimates of Analytic Quality Relative to Specified Quality Requirements for Glycohemoglobin and PSA*

Quality Requirement†	Group Mean	NTQ (σ)	NMQ (σ)	LMQ (σ)
Glycohemoglobin				
10%	9.06	1.93	1.93	2.57
1.0 %Hb	9.06	2.12	2.20	2.85
15%	9.06	2.89	3.21	3.86
20%	9.06	3.86	4.50	5.15
2.0 %Hb	9.06	4.24	5.04	5.69
25%	9.06	4.82	5.79	6.44
2.5 %Hb	9.06	5.30	6.47	7.11
PSA (%)				
10	6.5	1.17	0.87	1.76
20	6.5	2.34	2.63	3.52
30	6.5	3.51	4.39	5.28
40	6.5	4.67	6.15	7.04
50	6.5	5.84	7.91	8.80
10	19.0	1.17	0.86	1.68
20	19.0	2.34	2.54	3.35
30	19.0	3.50	4.22	5.03
40	19.0	4.67	5.89	6.71
50	19.0	5.84	7.57	8.39

Hb, hemoglobin; LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; PSA, prostate-specific antigen.
* Data from 5,066 participants for glycohemoglobin and 2,353 participants for PSA (College of American Pathologists specimens K-16 and K-17). The group means for glycohemoglobin are given as the percentage of hemoglobin; for PSA, in nanograms per milliliter.
† Quality requirements for glycohemoglobin are given in either allowable total error in units of percent (allowable total error in units of %Hb divided by a medical decision concentration of 7.0 %Hb) or actual concentration units of %Hb. The 2 sets of data for PSA represent different concentrations, the lower (6.5 ng/mL) applicable for screening applications, the higher (19.0 ng/mL) applicable for following treatment.

uniform national cutoffs to be applicable with PSA, today's methods do not provide reliable results within less than about 40% in the diagnostic range (as shown by the data for the 6.5 ng/mL PT specimen). Likewise, for elevated PSA values, values can only be assumed correct within 30% to 40% for today's methods (as shown by the results for the 19 mg/mL PT specimen). **Figure 6** shows the difficulty posed by current medical practices and current QC practices. Test results are not reliable for the current medical use in diagnosis and can be predicted to cause many false-positive results.

Table 3 shows the estimates of quality for prothrombin time (TE_a = 15%), INR (TE_a = 20%), and fibrinogen (TE_a = 20%), in which the quality requirements are defined by CLIA or CAP. The data show the results for 10 specimens and 2 PT events. The CAP survey did not provide an estimate of the group SD for prothrombin time or INR, thus NTQ could not be calculated and only estimates of NMQ and LMQ are shown in **Figure 7**, **Figure 8**, and **Figure 9**.

Discussion

These results show that analytic quality is still a major issue when evaluated on the σ scale. Three different σ metrics were calculated to account for measurement accuracy or bias in different ways. In the calculation of NTQ, the observed SD or CV of the group reflects method bias and imprecision, which is the simplest calculation and the most demanding estimate of quality. With the NMQ, the bias of each method subgroup is determined to calculate the σ for each subgroup, and then all method subgroups are summarized using a weighted average for the survey program. With the LMQ, bias is not considered, thus LMQ represents the most optimistic estimate and provides the highest σ values. For LMQ values to be applicable, test results must be interpreted against local method norms, local reference values, and local cutoffs to compensate for method bias in a particular laboratory; given that many laboratories lack the resources for such reference value studies, the LMQ values probably are representative of the performance available in very large clinical and reference laboratories. Despite the different calculations, all 3 estimates of quality led to the same conclusion—analytic quality is still a problem in US laboratories. The observed σ values generally are less than 4 and, in some cases, less than 3, a benchmark for minimal acceptable quality.

These estimates of quality are based on PT events during 2004. Results of PT testing were obtained from 5 PT programs that represent test applications ranging from physician office laboratories to large hospital laboratories. Participation in PT is required by CLIA, and all PT results are reported to CMS. Records from CMS show that approximately 36,000 laboratories participate in PT; however, not all laboratories participate for all tests. For example, during the fourth quarter of 2004, the number of laboratories participating in PT was 11,006 for cholesterol, 11,201 for calcium, and 11,784 for glucose. Our samples (Table 1) represent 84.1%, 87.4%, and

DOI: 10.1309/V50H4FRVVWX12C79

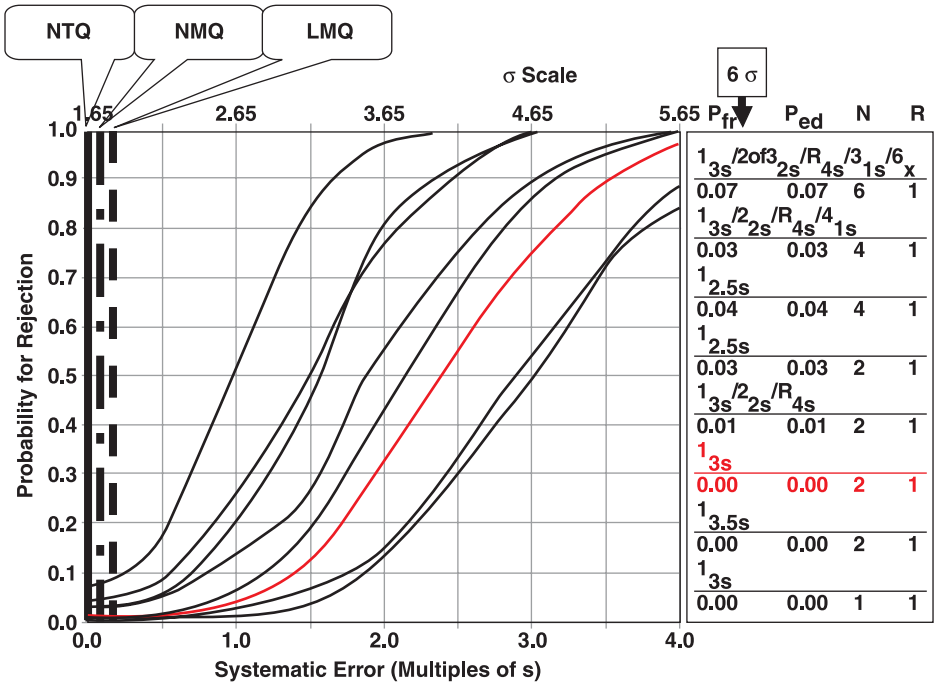

Figure 6 Prostatic-specific antigen quality in US laboratories for TE_a = 10%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

91.0% of those laboratories; thus the figures reported herein for cholesterol, calcium, and glucose should be representative of laboratory performance in the United States. We expect that the figures for glycohemoglobin should similarly represent a very large percentage of US laboratories. For PSA, prothrombin time, INR, and fibrinogen, which depend on data from CAP surveys, we expect that the performance figures actually are optimistic because CAP laboratories are seen to provide the best performance of the different survey programs.

The tests considered herein represent long-established and widely used laboratory tests. If we can generalize from the results for the 8 tests studied, the typical quality of a laboratory test is only 3 to 4 σ at best. On the σ scale, this level of quality is adequate only if the testing processes are controlled with much more QC than required by the CLIA minimum of 2 levels per day. Laboratories typically should be analyzing 4 or more control samples and interpreting the results with multirule procedures to maximize error detection.

Given the outcome of the present study, there will be concerns about the method used. First, the CLIA criteria for acceptability might not be objective and indicative of the quality needed for medical care; however, many clinical laboratory scientists think those criteria are too loose, not too tight. The quality criteria could be related to the actual medical use of the test results by considering the gray zone or clinical decision interval used in medical interpretation of critical test results and then deriving analytic total error criteria that are consistent with medical usefulness.[15,16] Nevertheless, the CLIA criteria are the established requirements for acceptability, right or wrong, and today's methods do not provide the performance necessary to guarantee that those criteria are satisfied in everyday operations. Second, not all of the tests considered herein are regulated analytes; thus, some do not have a national requirement for quality defined by the CLIA regulations. Glycohemoglobin and PSA are good examples, and one might wonder why these 2 tests are not on the regulated list because they are so important in health care. Our evaluation used a sensitivity analysis to demonstrate the available quality for a wide range of possible requirements and to identify the quality requirement that can be achieved at 5 to 6 σ.

There also might be concerns about the commutability of PT survey specimens[17] because they are not the same as real patient specimens. However, these specimens are the best specimens available to test the performance of laboratories across the nation. Professional organizations such as CAP continue to study the characteristics of test specimens,[18,19] and

DOI: 10.1309/V50H4FRVVWX12C79

Table 3
Estimates of Analytic Quality for Prothrombin Time, INR, and Fibrinogen on the Basis of 10 Specimens in the 2004 CAP Survey of 800 to 900 Laboratories

CAP Specimen	Mean*	NTQ (σ)†	NMQ (σ)	LMQ (σ)
Prothrombin time with TE_a = 15%				
CG2-01	12.3	—	3.07	6.38
CG2-02	11.6	—	2.33	5.18
CG2-03	12.6	—	2.94	5.75
CG2-04	14.0	—	3.03	5.93
CG2-05P	12.3	—	2.72	5.69
CG2-11	23.8	—	0.00	4.40
CG2-12	23.3	—	0.35	4.85
CG2-13	11.9	—	3.10	6.47
CG2-14	26.0	—	0.00	4.17
CG2-15P	20.6	—	0.92	4.66
Summary	**16.8**	—	**1.77**	**5.35**
INR with TE_a = 20%				
CG2-01	1.00	—	2.57	3.59
CG2-02	0.94	—	2.63	3.19
CG2-03	1.05	—	2.54	3.51
CG2-04	1.20	—	2.19	3.24
CG2-05P	1.02	—	2.74	3.47
CG2-11	2.41	—	2.10	3.69
CG2-12	2.39	—	1.91	3.53
CG2-13	0.97	—	3.04	3.91
CG2-14	2.73	—	2.06	3.58
CG2-15P	2.03	—	2.13	3.48
Summary	**1.57**	—	**2.39**	**3.52**
Fibrinogen with TE_a = 20%				
CG2-01	249	2.41	2.59	3.50
CG2-02	294	2.11	2.33	3.25
CG2-03	207	1.41	1.37	2.97
CG2-04	168	1.04	0.66	2.56
CG2-05P	205	1.42	1.80	3.47
CG2-11	352	2.02	2.44	3.32
CG2-12	237	1.54	2.06	3.40
CG2-13	227	2.27	2.41	3.38
CG2-14	355	2.08	2.57	3.45
CG2-15P	204	1.54	1.84	3.14
Summary	**260**	**1.78**	**2.01**	**3.24**

CAP, College of American Pathologists; INR, international normalized ratio; LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; TE_a, allowable total errors.
* The means for prothrombin time are given in seconds; for fibrinogen, in milligrams per deciliter. Système International (SI) units also are seconds for prothrombin time; to convert fibrinogen values to SI units (g/L), multiply by 0.01. Summary figures, in bold, are weighted averages that account for the relative number of laboratories in the respective PT groups.
† Metrics for NTQ could not be calculated because the group SD was not reported.

ongoing improvements should be a high priority when such data are required in the regulatory process. On the other hand, the results on PT specimens actually might be overly optimistic because laboratories pay special attention to the analysis of these specimens. PT specimens often are analyzed right after calibration and often are bracketed by control materials; thus, the results might represent the optimal "in-control" performance of laboratories. More reliable estimates might be available from peer-comparison programs, in which laboratories submit all QC data for analysis and comparison with the peer group; however, those data are not readily available except to the providers and participants in the particular peer program.

Despite these possible limitations, this evidence of poor quality laboratory testing should be taken seriously because it is the most quantitative assessment available. Furthermore, the observed results are entirely understandable and predictable based on past practices in laboratory medicine and evolving methods in evidence-based medicine.

Consider past practices to improve the quality of lipid testing. Beginning with the National Cholesterol Education Program (NCEP) guidelines in the late 1980s,[20] desirable precision was specified as a CV of 3% or less and desirable accuracy as a bias of 3% or less. Then in 1992, CLIA defined an allowable total error of 10%. Given the combined NCEP and CLIA guidelines, the quality that would be expected would be 2.33 σ [(10 – 3)/3] to 3.33 σ [10/3] if bias were zero. Later NCEP guidelines for other lipid tests[21-23] formalized a goal-setting method in which the allowable bias plus 2 times the allowable SD or CV are set equal to the allowable total error, eg, triglycerides specifications were set as an allowable CV of

DOI: 10.1309/V50H4FRVVWX12C79

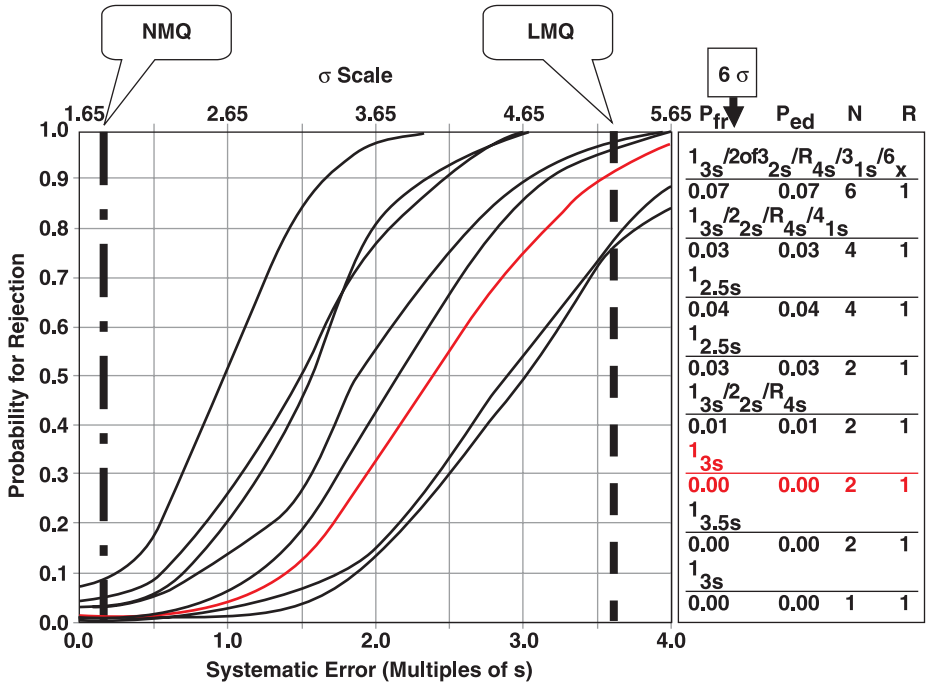

Figure 7 Prothrombin time quality in US laboratories for TE_a = 15%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; s, standard deviation; TE_a, allowable total errors.

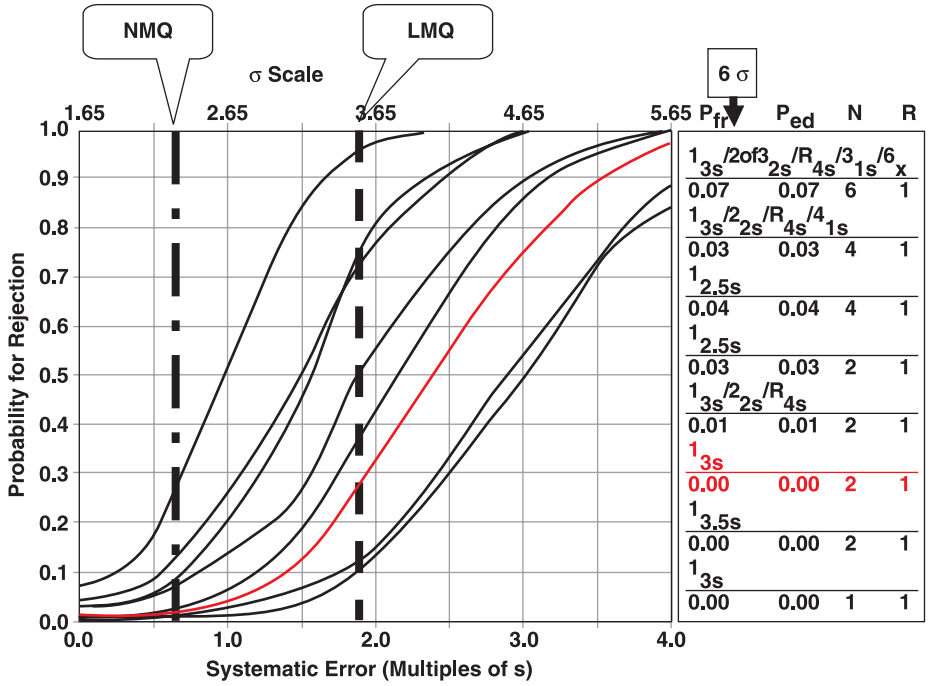

Figure 8 International normalized ratio quality in US laboratories for TE_a = 20%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; s, standard deviation; TE_a, allowable total errors.

DOI: 10.1309/V50H4FRVVWX12C79

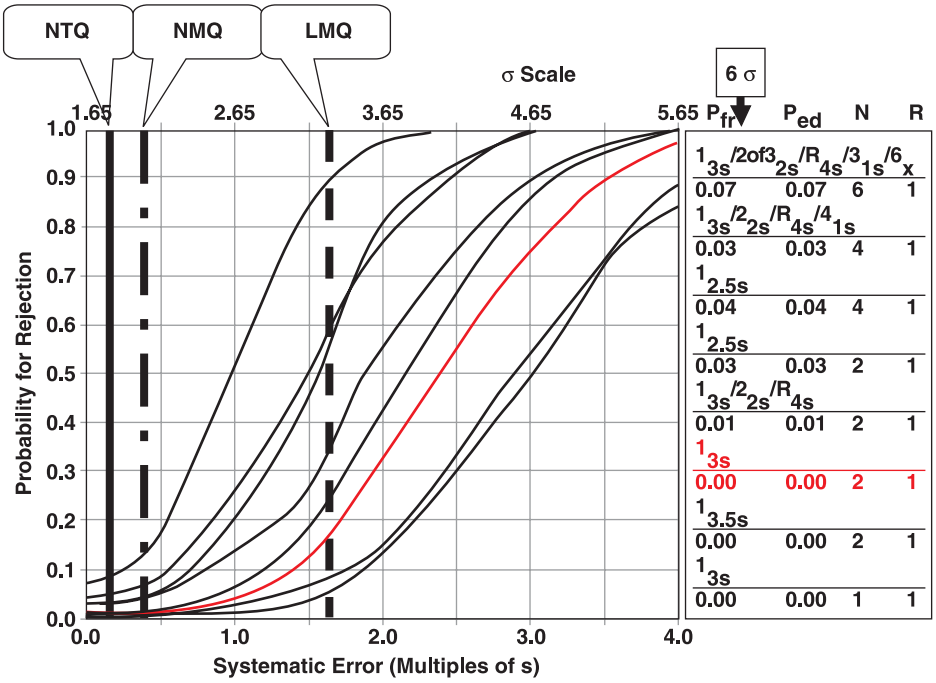

Figure 9 Fibrinogen quality in US laboratories for TE_a = 20%. The curves represent the statistical power of the different quality procedures whose control rules and total number of control measurements (N) are listed in the key at the right. P_{fr} is the probability for false rejections, which is determined from the y-intercept of the power curve. P_{ed} should be determined at the intersection of the vertical lines and the power curves. R represents the number of runs over which the control rules are applied, which is a single run for all the QC procedures shown here. Solid line, NTQ; dash-dot line, NMQ; dashed line, LMQ. LMQ, local method quality; NMQ, national method quality; NTQ, national test quality; s, standard deviation; TE_a, allowable total errors.

5%, allowable bias of 5%, and allowable total error of 15%; high-density lipoprotein cholesterol specifications were set as an allowable CV of 6%, allowable bias of 10%, and allowable total error of 22%; low-density lipoprotein cholesterol specifications were set as an allowable CV of 4%, allowable bias of 4%, and allowable total error of 12%. Thus, past practices for planning measurement procedures have aimed at 2 to 3 σ quality, which is the quality observed in the present study.

Consider also the evolving method of evidence-based medicine. There is not yet any scientific model for setting method specifications on the basis of the medical use of the test. Glycohemoglobin provides a good example in which medical usefulness or clinical quality depends on being able to distinguish a test result of 8.0 %Hb from a true value of 7.0 %Hb to initiate treatment.[24] Given the known within-subject biologic variation of 4.1%[25] and the specified maximum CV of 5% and assuming a bias of 0.0%, this testing process is not controllable to the quality required for patient care.[26] If the desirable CV of 3% were achieved (again with bias of 0.0%), the testing process would be controllable but would require 4 to 6 control measurements per run with multirule interpretation. The assumption of zero bias is certainly not realistic; thus, it should be expected that quality will be only 2 to 3 σ as confirmed in the present study (Table 2, quality requirement of 1.0 %Hb, σ values from 2.12 to 2.85).

Conclusions

Given such poor quality for laboratory tests when evaluated on the σ scale, it makes no sense to reduce daily QC to weekly or even monthly, as proposed in the SOM guidelines for EQC procedures. In fact, the results of the present study suggest that QC generally should be increased to at least 4 controls per run and that control rules should be selected to maximize error detection. For optimal testing, laboratories should design their QC procedures for each individual test to account for the precision and accuracy of their measurement procedures and the quality required for care of their patients.[27] CLIA's minimum QC of 2 levels per day should apply only to measurement procedures that demonstrate 5 σ quality or higher. For any measurement procedure to be eligible for EQC procedures, it should be required to demonstrate 6 σ quality. The application of Six Sigma principles and metrics would greatly improve the proposed EQC validation process and provide a scientific basis for recommendations on the amount of QC that is needed.[28]

DOI: 10.1309/V50H4FRVVWX12C79

From the [1]Department of Pathology and Laboratory Medicine, University of Wisconsin Medical School; and [2]Westgard QC, Madison.

Address reprint requests to Dr Westgard: Department of Pathology and Laboratory Medicine, University of Wisconsin Medical School, Madison, WI 53706.

References

1. Ross JW, Boone DJ. Assessing the effect of mistakes in the total testing process on the quality of patient care [abstract]. *Proceedings of the 1989 Institute on Critical Issues in Health Laboratory Practice*. Atlanta, GA: Centers for Disease Control; 1991:173.
2. Plebani M, Carraro P. Mistakes in a state laboratory: types and frequency. *Clin Chem*. 1997;43:1348-1351.
3. Bonini P, Plebani M, Ceriotti F, et al. Errors in laboratory medicine. *Clin Chem*. 2001;48:691-676.
4. Downer K. Five years after "*To Err is Human*." *Clin Lab News*. February 2005;31:1, 6, 7.
5. Centers for Disease Control and Prevention (CDC), Centers for Medicare & Medicaid Services (CMS), HHS. Medicare, Medicaid, and CLIA programs: laboratory requirements relating to quality systems and certain personnel qualifications: final rule. *Fed Regist*. 2003;68:3640-3714.
6. *CMS State Operations Manual: Interpretive Guidelines*. Baltimore, MD: Centers for Medicare and Medicaid Services; 2004.
7. Westgard JO: Hear, hear, hear! Hearings on untruth and unequality! Part V: a few bad apples or the tip of the iceberg? Available at www.westgard.com/essay68.htm. Accessed January 10, 2006.
8. Harry M, Schroeder R. *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations*. New York, NY: Currency; 2000.
9. Nevalainen D, Berte L, Kraft C, et al. Evaluating laboratory performance on quality indicators with the Six Sigma scale. *Arch Pathol Lab Med*. 2000;124:516-519.
10. Westgard JO. *Six Sigma Quality Design & Control: Desirable Precision and Requisite QC for Laboratory Measurement Processes*. Madison, WI: Westgard QC; 2000.
11. Westgard JO, Burnett RW. Precision requirements for cost-effective operation of analytical processes. *Clin Chem*. 1990;36:1629-1632.
12. American Society for Quality Control, Chemical and Process Industries Division, Chemical Interest Committee. *Quality Assurance for the Chemical and Process Industries: A Manual of Good Practices*. Milwaukee, WI: ASQC Quality Press; 1986:37.
13. Chesher D, Burnett L. Equivalence of critical error calculations and process capability index Cpk. *Clin Chem*. 1997;43:1100-1101.
14. Petersen PH, Ricos C, Stockl D, et al. Proposed guidelines for the internal quality control of analytical results in the medical laboratory. *Eur J Clin Chem Clin Biochem*. 1996;34:983-999.
15. Westgard JO, Wiebe DA. Cholesterol operational process specifications for assuring quality required by CLIA proficiency testing. *Clin Chem*. 1991;37:1938-1944.
16. Westgard JO, Petersen PH, Wiebe DA. Laboratory process specifications for assuring quality in the US National Cholesterol Education Program. *Clin Chem*. 1991;37:656-661.
17. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. *Clin Chim Acta*. 2003;327:25-37.
18. Challenging the proficiency testing process: lessons from CAP's 2003 fresh frozen serum project. *Clinical Laboratory Strategies*. August 2004;9:1, 8-9.
19. Miller WG, Myers GL, Ashwood ER, et al. Creatinine measurement: state of the art in accuracy and interlaboratory harmonization. *Arch Pathol Lab Med*. 2005;129:297-304.
20. National Cholesterol Education Program Laboratory Standardization Panel. Current status of blood cholesterol measurement in clinical laboratories in the United States. *Clin Chem*. 1988;34:193-201.
21. Bachorik PS, Ross JW, for the NCEP Working Group on Lipoprotein Measurement. National Cholesterol Education Program recommendations for measurement of low-density lipoprotein cholesterol: executive summary. *Clin Chem*. 1995;41:1414-1420.
22. Stein EA, Myers GL, for the NCEP Working Group on Lipoprotein Measurement. National Cholesterol Education Program recommendations for triglyceride measurement: executive summary. *Clin Chem*. 1995;41:1421-1426.
23. Warnick GR, Wood PD, for the NCEP Working Group on Lipoprotein Measurement. National Cholesterol Education Program recommendations for measurement of high-density lipoprotein cholesterol: executive summary. *Clin Chem*. 1995;41:1427-1433.
24. Sacks DB, Bruns DE, Goldstein DE, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clin Chem*. 2002;48:436-472.
25. Larsen ML, Fraser CG, Petersen PH. A comparison of analytical goals for haemoglobin A_{1c} assays derived using different strategies. *Ann Clin Biochem*. 1991;28(pt 3):272-278.
26. Westgard JO. Clinical quality vs analytical performance: what are the right targets and target values? *Accreditation and Quality Assurance*. 2004;10:10-14.
27. Westgard JO. Internal quality control: planning and implementation strategies. *Ann Clin Biochem*. 2003;40:593-611.
28. Westgard JO, Westgard SA. Equivalent quality testing versus equivalent QC procedures. *Lab Med*. 2005;36:726-729.

DOI: 10.1309/V50H4FRVVWX12C79