Clinical Chemistry 56:5
714–722 (2010)

Reviews

Establishment of Outcome-Related Analytic Performance Goals

George G. Klee[1*]

BACKGROUND: Accrediting organizations require laboratories to establish analytic performance criteria that ensure their tests provide results of the high quality required for patient care. However, the procedures for instituting performance criteria that are directly linked to the needs of medical practice are not well established, and therefore alternative strategies often are used to create and implement surrogate performance standards.

CONTENT: We reviewed 6 approaches for establishing outcome-related analytic performance goals: (*a*) limits defined by regulations and external assessment programs, (*b*) limits based on biologic variation, (*c*) limits based on surveys of clinicians about their needs, (*d*) limits based on effects on guideline driven medical decisions, (*e*) limits based on analysis of patterns for ordering follow-up clinical tests, and (*f*) limits based on formal medical decision models. Performance criteria were tabulated for 12 common chemistry analytes and 4 routine hematology tests.

CONCLUSIONS: There is no consensus currently about the preferred methods for establishing medically necessary analytic performance limits. The various methods reviewed give considerably different performance limits. The analytic performance limits claimed by a laboratory should correspond to those limits that can be reliably maintained based on validated QC monitoring systems. These limits generally are larger than the observed CVs and bias parameters collected for assay validation. There is a major need for increased communication among laboratorians and clinicians on this topic, especially when the analytic performance limits that can be consistently maintained by a laboratory are inconsistent with the expectations of health care providers.

[1] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN.
* Address correspondence to the author at: Mayo Clinic, 200 First St. S.W., Rochester, MN 55905. Fax 507-266-5193; e-mail klee.george@mayo.edu.
Received November 20, 2009; accepted March 1, 2010.
Previously published online at DOI: 10.1373/clinchem.2009.133660

Objective performance limits are fundamental requirements for the evaluation and effective control of laboratory systems. The CLIA Amendments (CLIA-88) require that the laboratory director "ensure that the test methodologies selected have the capability of providing the quality of results required for patient care and ensure the establishment and maintenance of acceptable levels of analytic performance for each test system" (CFR sec 493.1445) *(1)*. Also, clearly articulated analytic performance criteria are needed for the development of robust QC systems because they are essential for calculating false-positive and false-negative detection rates *(2)*. However, the procedures for establishing analytic performance specifications that are required for an assay to meet defined clinical utility are not well prescribed.

Fraser published a hierarchical approach to classification of strategies to set quality specifications in laboratory medicine *(3)*. This hierarchical list of strategies was endorsed by an international conference, Strategies to Set Global Quality Specifications in Laboratory Medicine, and has been termed "The Stockholm Conference Hierarchy" *(4, 5)*. Fraser's top 4 strategies were: (*a*) assessment of effect on clinical decision making, (*b*) professional recommendations from national and international expert groups and expert individuals and institutional groups, (*c*) regulatory and external assessment specifications, and (*d*) published data on the state of the art.

Fraser noted that although specifications based on how quality affects medical decision making are at the top of the hierarchy, this approach is difficult to apply because few tests are used in single well-defined clinical situations with standardized widely accepted medical strategies that are directly related to the test results. He also noted that the analysis of the effects of assay performance on medical decisions is heavily dependent on the assumptions made about how the test results are used by the clinicians. Therefore this first strategy is seldom used.

A recent College of American Pathologists Q-Probe analysis of physician satisfaction with clinical laboratory services showed that the category most frequently selected as the most important was "quality/reliability of results" *(6)*. It is interesting to note that this category, which encompassed the trueness and precision of test results, had one of the highest levels of

satisfaction, whereas other categories, such as turnaround times, adequacy of test menu, and courier services, had lower levels of satisfaction. Perhaps it is easier for clinicians to objectively evaluate these other categories, which have well-defined performance expectations, than to evaluate quality/reliability of the test results, which do not have well-defined performance expectations. In the absence of objective analytic performance limits defining the "quality of results required for patient care," it also is difficult for laboratorians to assure that their QC systems are reliably detecting inadequate performance. Even when a laboratory maintains well-defined analytic performance specifications, Plebani has cogently noted that this may have little effect on clinical decision making unless the clinicians are well informed about these specifications *(7)*.

This report reviews some of the published methods for establishing assay performance specifications and discusses the strengths and weaknesses of these approaches and their implications for enhancing medical care. This review shows that there is not a single strategy for establishing outcome-related analytic performance goals, and multiple interrelated approaches may be necessary. A decade ago Werner noted that the linkage between analytic goals and medical care strategies is reciprocal in that clinical outcome can be optimized by either tailoring medical strategy to existing analytic performance or optimizing analytic performance to meet medical strategies *(8)*. Six approaches have been proposed that can serve as a basis for establishing assay performance limits: regulations and external assessments, biologic variation, surveys of clinician needs, guideline-driven decisions, analysis of clinical ordering patterns, and formal decision models.

Approach 1: Performance Limits Defined by Regulations and External Assessment Specifications

Regulatory agencies and external proficiency-testing programs have established performance limits for evaluating interlaboratory proficiency testing (PT).[2] Those limits are based predominately on the state of the art and/or biologic variability. These limits are used as surrogates for clinical performance because well-defined clinical performance limits generally are not available *(9)*.

CLIA-88 defined interlaboratory PT precision limits for many routine assays. Although the procedures used to develop these PT limits are not explicitly provided; the limits are implicitly linked to the state-of-the-art laboratory practice as of 1988. The limits enacted by CLIA-88 caused only a few laboratories to fail the PT program when these limits were used to grade the surveys. PT performance limits for selected chemical laboratory tests are shown in Table 1. Some of these limits are quite wide. For example, the limits for calcium are defined as ±10 mg/L. These limits span an interval wider than the reference interval. Differences in serum calcium of smaller magnitudes could have major clinical implications.

The German External Quality Assessment Scheme (G-EQAS) uses the concept of root mean square deviation (RMSD) to define performance limits *(10)*. The RMSD is based on trimming 10% of the data and calculating a percentage error relative to a target value *(10)*. Some external assessment groups, such as the external quality assessment program of the Centre of Biomedical Research (Treviso, Italy) have established performance limits derived from total analytic error (TE_a) *(11, 12)*. These total error limits combine imprecision and inaccuracy limits to provide a combined statistic derived from biologic variation for within-subject (CV_I) and between-subject (CV_G) CVs: (TE_a = imprecision + inaccuracy, also called bias). As discussed in the following section, a goal for desirable imprecision is $<0.5\ CV_I$, and a goal for bias is $<0.25\ \sqrt{(CV_I^2 + CV_G^2)}$. The specific performance limits generally include a coverage factor (k), which indicates how rigorously the limits are applied. A k factor of 1.65 corresponds to imprecision coverage to the 95th percentile, whereas a k factor of 1.96 corresponds to imprecision coverage up to the 97.5th percentile. In other words, with a k=1.65, 5% of the results may exceed the upper performance limit, and with a k=1.96, only 2.5% of the results may be expected to exceed the upper limit. Examples of some of performance limits for CLIA, G-EQUAS, and TE_a are shown in Table 1. For this table, total error limits are calculated as: $k \times (0.5\ CV_I) + 0.25\ \sqrt{(CV_I^2 + CV_G^2)}$, where k = 1.65.

Neither the CLIA-88 nor the external quality assessment systems have absolute limits for evaluation of trueness. Both systems use statistical methods to assign target values that are peer-group specific. The implicit assumption is that clinicians compensate for the between-method differences in assay distribution based on the reference ranges and interpretative information provided by the laboratories, but little evidence is available to validate this assumption.

[2] Nonstandard abbreviations: PT, proficiency testing; G-EQAS, German External Quality Assessment Scheme; RMSD, root mean square deviation; TE_a, total analytic error; CV_I, within-subject CV; CV_G, between-subject CV; k, coverage factor; TSH, thyroid-stimulating hormone; PSA, prostate-specific antigen; PTH, parathyroid hormone.

Table 1. Performance limits based on CLIA[a] and G-EQUAS[b] proficiency testing limits and biologic variation.

		Based on proficiency testing		Based on biologic variation[c]				
Analyte	Units	CLIA limits[a]	G-EQUAS, %RMSD[b]	CV_I	CV_G	Imprecision	Bias	Total error
Bilirubin	mg/L	±4 (or 20%)	13.0 (>20g/L)	23.8%	39.0%	11.9%	11.4%	31.1%
Calcium	mg/L	±10	6.0	1.9%	2.8%	0.8%	0.8%	2.4%
Cholesterol	mg/L	±10%	7.0	5.4%	15.2%	4.0%	4.0%	8.5%
Cortisol	μg/L	±25%	16.0	20.9%	45.6%	12.5%	12.5%	29.8%
Creatinine	mg/L	±3 (or 15%)	11.5	5.3%	14.2%	3.8%	3.8%	8.2%
Glucose	mg/L	±60 (or 10%)	11.0	5.7%	6.9%	2.9%	2.2%	6.9%
Iron	μg/L	±20%		26.5%	23.2%	13.3%	8.8%	30.7%
Phosphorus	mg/L	±0.3 (or 10.7%)	9.0	8.5%	9.4%	4.3%	3.2%	10.2%
Potassium	mmol/L	±0.5	4.5	4.8%	5.6%	2.4%	1.8%	5.8%
Sodium	mmol/L	±4.0	3.0	0.7%	1.0%	0.4%	0.3%	0.9%
Thyroxine	μg/L	±10 (or 20%)	12.5	4.9%	10.9%	2.5%	3.0%	7.0%
Total protein	g/L	±10%	6.0	2.7%	4.0%	1.4%	1.2%	3.4%
Triglycerides	mg/L	±25%	9.0	20.9%	37.2%	10.5%	10.7%	27.9%
Hematocrit	%	6%	5.0	2.8%	6.4%	1.4%	1.7%	4.1%
Hemoglobin	g/L	7%	4.0	2.8%	6.6%	1.4%	1.8%	4.1%
Leucocytes	10^9/L	15%	6.5	10.9%	19.6%	5.6%	5.6%	14.6%
Erythrocyte mean cell volume	fL	±3SD		1.3%	4.8%	0.7%	1.2%	2.3%

[a] CDC *(1)*.
[b] Westgard *(10)*.
[c] http://www.westgard.com/guest17.htm.

Approach 2: Using Biologic Variation to Formulate Boundaries for Analytic Variation Limits

Statistically it can be shown that the total variation of test results is the combination of the analytic precision (expressed as SD or CV) and the biologic variations (also expressed as SD or CV). The total system variation (assuming independence) is the square root of the sum of the squared analytic and biologic SD:

$$SD_{total} = \sqrt{(SD^2_{analytic} + SD^2_{biologic})}.$$

The choice of an individual biologic CV_I vs a group biologic CV_G depends on how the test is used in clinical practice. For monitoring changes in individual patients over time, CV_I is most appropriate. For classifying patients into diagnostic or therapeutic categories using reference intervals or action limits, CV_G is more important. It should be noted that most of the estimates for CV_I and CV_G are derived from healthy individuals, whereas most medical decisions are made in sick patients. In patients, diseases often lead to different concentrations of the target analytes and different magnitudes of biologic variability.

Statistically, biologic variation can be used to formulate boundaries for the allowable limits for analytic precision, because if $SD_{analytic}$ is small compared to $SD_{biologic}$, only minor increments to the total variation are added. The following terminology has been proposed *(11, 12)*:

Minimal performance $CV_{analytic} < 0.75\ CV_I$ (adds <25% to CV_{total});

Desirable performance $CV_{analytic} < 0.50\ CV_I$ (adds <12% to CV_{total});

Optimal performance $CV_{analytic} < 0.25\ CV_I$ (adds <3% to CV_{total}).

Subtle issues influence this approach, for example, the effects of preanalytic variables on biologic variation. If preanalytic collection variables are controlled, for example, time of day, position, diet control, time of tourniquet application, time of drug administration, and source (whether samples are from ambulatory or hospitalized patients), then the biologic variation is likely to be smaller. On the other hand, patients with altered pathophysiology are likely to have larger biologic variation than healthy individuals.

An added drawback of this approach occurs when these imprecision performance limits are expanded to provide total error limits. Various authors have proposed that performance limits for analytic bias also can be statistically derived from the biologic variations. Gowans et al. proposed that for laboratories to share common reference intervals, analytic bias should be $<0.25\sqrt{(CV_I^2 + CV_G^2)}$ *(13)*. Fraser and Peterson subsequently recommended the following terminology for analytic bias, which is analogous to the terminology recommended for imprecision *(14)*:

Minimum bias performance $<0.375\sqrt{(CV_I^2 + CV_G^2)}$;

Desirable bias performance $<0.250\sqrt{(CV_I^2 + CV_G^2)}$;

Optimum bias performance $<0.125\sqrt{(CV_I^2 + CV_G^2)}$.

The biologic variations and the "desirable" performance limits for selected analytes based on the information compiled by Ricos et al. *(15)* are summarized in Table 1.

A major concern about these bias goals is that they are targeted at decisions for individual patients who are subject to biologic variations; however, analytic bias produces a systemic shift of all test results that can produce major changes in the medical decisions for a large number of patients. Biologic variation may broaden the distribution of test values in this aggregate cohort, but it generally does not shift the center point of the distribution, whereas analytic bias directly shifts the position of the distribution. Analytic bias can have a profound effect near clinical decision points. This concept is discussed further in approach 4, which addresses analytic performance characteristics based their effects on guideline-driven medical decisions.

Approach 3: Surveys of Clinician Opinion of Tolerable Laboratory Changes

Various studies have estimated the magnitude of test-value differences that would cause clinicians to alter their patient care plans. Generally these studies have presented specific case studies with various levels of selected analytes. The limits for analytic variations tolerated by clinicians are dependent on the experience of the clinicians, with the more experienced clinicians having greater tolerance for variations. A widely referenced study about the medical significance of laboratory tests was conducted by Barnett in the 1960s *(16)*. He used "expert opinion" to estimate how laboratory test variation affected medical decisions. Similarly, in the 1970s, Skendzel published a report in JAMA based on a survey of 125 internists using a series of case studies followed by a list of alternate test values to ascertain what magnitude of test value change would cause changes in the clinician's decisions *(17)*. Skendzel followed this up with a larger survey of AMA physicians that was used to define analytic performance limits necessary to meet medical utility *(18)*. More recently, Thue et al. surveyed Norwegian general practitioners and found an analytic imprecision limit for hemoglobin of 2.8% *(19)*. Similarly, Skeie et al. surveyed patients who performed self-monitoring of glucose and recommended imprecision limits of 3.1%–5.0% based on their decisions for insulin dose adjustments *(20)*.

Some of the performance limits are summarized in Table 2. Skendzel and colleagues coined a term called "medical coefficient of variation," which is calculated by using a statistical conversion factor to convert the maximum allowable change to something similar to an SD. The "medically significant" analyte difference is divided by 1.645 and $\sqrt{2}$ to convert it to an SD, then this number is converted to a CV by multiplying by 100 and dividing by the average of the original and maximum allowable test values. The values given in Table 2 may seem large and easy to maintain, but the analytic CV required for a laboratory to guarantee this level of performance at even a 3 to 5 sigma level, would be one-third to one-fourth of this performance level.

Because there are statistical methods for assessing biologic and analytic variation, perhaps those variance parameters could be better evaluated by using scientific protocols. However, expert opinion of clinicians may be very valuable for estimating the clinical impact of analytic bias and the impact of aberrant test results. Medical decisions generally are based on an aggregated set of observations and measurements for each patient. Assessment of inconsistencies across multiple factors is a valuable tool for identifying erroneous test values before they cause adverse medical complications.

Inquiries from experienced clinicians, especially those working in specialized practices, can be valuable early warnings for laboratory performance problems. For example, an abrupt increase in the frequency of laboratory reports with hypercalcemia may signify an upward shift in serum calcium measurements *(21)*. An increased discordance between thyrotropin and free thyroxine hormone measurement may signify analytic problems with 1 or the other of these tests *(22)*. Similarly, increased discordance between the erythrocyte mean cell volume and serum measurements of iron, ferritin, vitamin B_{12}, and/or folate could be early warnings of analytic problems *(23)*. Also, concerns of clinicians about the specificity and analytic detection limits of immunoassays for cortisol and testosterone may highlight unmet analyte performance expectations *(24–28)*.

Table 2. Performance limits and medical utility[a] based on physicians opinions and analytic bias based on population distributions.[b]

		Medical utility, % CV			Population analytic bias limits		
Analyte	Units	Base value	Change value	Medical CV[a]	Decision limit	Bias limit[b]	Bias, CV
Bilirubin	mg/L	8	14	23.4%	11	±1	9.0%
Calcium	mg/L	90	106	7.0%	102	±1	1.0%
Cholesterol	mg/L	2100	2800	12.3%	2000	±23	1.2%
Creatinine	mg/L	10	15	17.2%	8	±1	12.5%
Glucose	mg/L	100	130	11.2%	1000	±20	2.0%
Iron	μg/L	150	100	17.2%	—	—	—
Phosphate	mg/L	350	250	14.3%	25	±1	4.0%
Potassium	mmol/L	3.8	3.4	4.8%	3.6	±0.1	2.8%
Sodium	mmol/L	125	130	1.7%	134	±1.5	1.1%
Thyroxine	μg/L	60	40	17.2%	50	±4	8.0%
Total protein	g/L	70	85	8.3%	63	±2	3.2%
Triglycerides	mg/L	1300	1900	16.1%	4000	±58	1.5%
Hematocrit	%	42	37	5.4%	35	±0.7	2.0%
Hemoglobin	g/L	150	138	3.6%	119	±3	2.5%
Leucocytes	10^9/L	6.0	3.4	16.4%	3.5	±0.2	5.7%
Erythrocyte mean cell volume	fL	95	100	3.2%	81.5	±0.7	1.0%

[a] Medical CV = 100 × [(change value − base value)/(1.645 × √2)]/[(change value + base value)/2]).
[b] Bias limit = 1 SD of change of population cumulative frequency distribution.

Approach 4: Analytic Performance Characteristics Based on Their Effects on Guideline-Driven Medical Decisions

Strategies for standardizing medical practice generally are based on a combination of expert opinion and analyses of published reports of studies. Most of these "guidelines" have been developed by clinicians who assume that assay results from accredited laboratories are uniformly of high quality and harmonized. Details on standardization of assays and/or analytic performance limits seldom are included with the decision limits defined in these guidelines.

Analytic bias can have a profound effect on the percentage of patients included in each branch of a guideline. The impacts of analytic bias on 3 practice guidelines were evaluated: (*a*) use of serum cholesterol for identifying patients at risk for coronary artery disease, (*b*) use of serum thyrotropin for detecting primary hypothyroidism, and (*c*) use of serum prostate-specific antigen (PSA) in prostate cancer screening *(29)*. A 3% positive bias effect for cholesterol increased the number of patients at risk for coronary artery disease by 16.7% at a cholesterol concentration of 2000 mg/L. A 6% positive bias on thyroid-stimulating hormone (TSH) caused a 26.6% increase in patients screened for hypothyroidism at a decision concentration of 5.0 mIU/L. A 6% positive bias in PSA caused an 11.4% increase in the number of men whose screening results were positive for prostate cancer at a PSA cutoff of 4.0 μg/L. Similar percentage increases were found at alternate screening cutoffs for each of these tests. Although the best-practice goal for analytic bias would be zero bias, attaining this goal would be very expensive and generally not feasible. Analyses of the adverse impact of analytic bias on medical decisions, such as those illustrated above, may provide useful cost-benefit comparisons. The cost of reducing analytic bias can be related to the potential benefits of better practice. Shermock et al. studied the effects of laboratory test variation in the international normalization ratio based on a defined set of clinical actions driven by international normalization ratio results *(30)*. Shermock et al. found that the influence of analytic errors on medical decisions depended on how close a value was to the decision limits and cautioned against using the same acceptability limits for all test levels.

A potential procedure for establishing tolerance limits for analytic bias is based on modeling the short-term variations of distributions of test values in the patient population *(31)*. The cumulative frequency distribution for 20 consecutive data sets of approximately 1000 test values each (corresponding to patients seen at Mayo Clinic, Rochester each day) are shown in

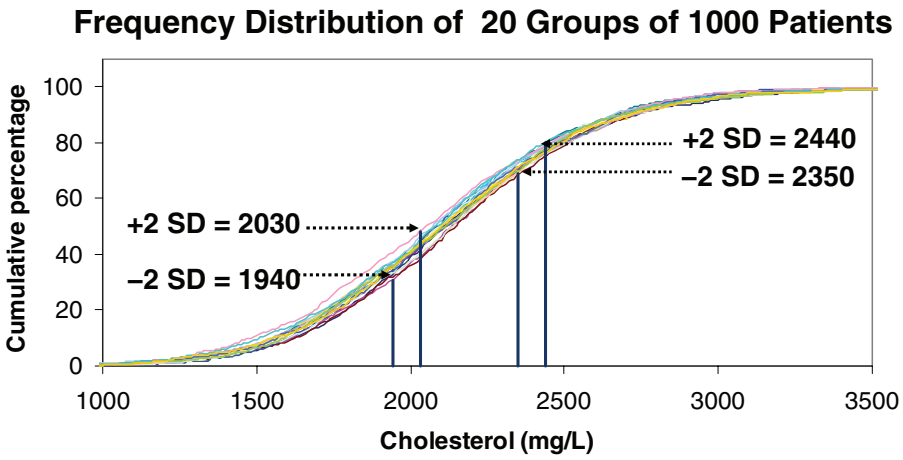

Fig. 1. Use of variations of the cumulative distributions of test values from large cohorts of patients to define analytic bias performance criteria.
The 2 SD limits for decision limits at 2000 and 2400 mg/L are illustrated. Analytic bias performance limits are set at 1 SD.

Fig. 1. The day-to-day population variation for practice guideline decisions can be considered similar to the within-individual biologic variation for individual patient decisions. By analogy, if the laboratory keeps the day-to-day analytic bias small compared to the shifts in the frequency distribution, the effects on the practice should be negligible. This model proposes that the analytic bias be kept at less than 1 SD of the population variation. In Fig. 1 the range between the −2 SD and +2 SD limits for cholesterol at both the 2000 mg/L and 2400 mg/L decision limits is 90 mg/L, so the 1 SD bias tolerance limit is 23 mg/L. The bias tolerance parameters for cholesterol and other chemistry and hematology measurements calculated using this model are shown in Table 2 *(31, 32)*. These limits are relatively small, but they illustrate how these small changes in analytic bias can directly affect patient care.

Approach 5: Analytic Performance Characteristics Based on Relationship to Clinical Procedure Orders

Clinical decisions are based on many factors, including the patients presenting problems, previous history, family history, the results of laboratory tests or procedures, and the preferences of the health care providers. In a comprehensive health care facility there will be considerable differences in the ways that laboratory tests are used, so it is difficult to analyze the exact relationship between test results and medical actions. For example, the serum parathyroid hormone assay (PTH) is a logical follow-up procedure to be ordered in a patient with new-onset hypercalcemia. If all health care providers responded in the same manner, one would expect an increase in the frequency of PTH test requests as a function of increases in calcium concentration above the upper limit of the reference interval. In the care of individual patients, however, there is considerable variation in the ordering patterns for PTH tests, due to issues of patient differences and provider preferences. On the other hand, in analysis of the care of large numbers of patients, there is a defined relationship between the concentrations of serum calcium and the relative frequency of PTH assays ordered within a short time after the calcium value is reported *(33)*.

The relationship between test values and the frequency of follow-up procedures for a particular medical center can be determined by combining 2 sets of data that are frequently computerized: the laboratory reports and the billing procedure codes. Because the relationship between these variables often is nonlinear, mathematical curve-fitting programs may be needed to define these relationships. In the calcium example depicted in Fig. 2, a total of about 100 000 serum calcium reports were sorted into 4 categories (male and female with Medicare and non-Medicare billing). For each category the numbers of patients within each specific 1 mg/L value of calcium concentration were enumerated, and the relative frequencies of the follow-up CPT4 codes (shown in parenthesis) were calculated. Many CPT4 codes in addition to PTH (83 970) were found to have ordering patterns associated with serum calcium concentrations, including requests for additional serum calcium (82 310), urine calcium (82 340), serum alkaline phosphate (84 075), chest x-ray (71 020), and nuclear scan of the parathyroid (78 070) *(33)*.

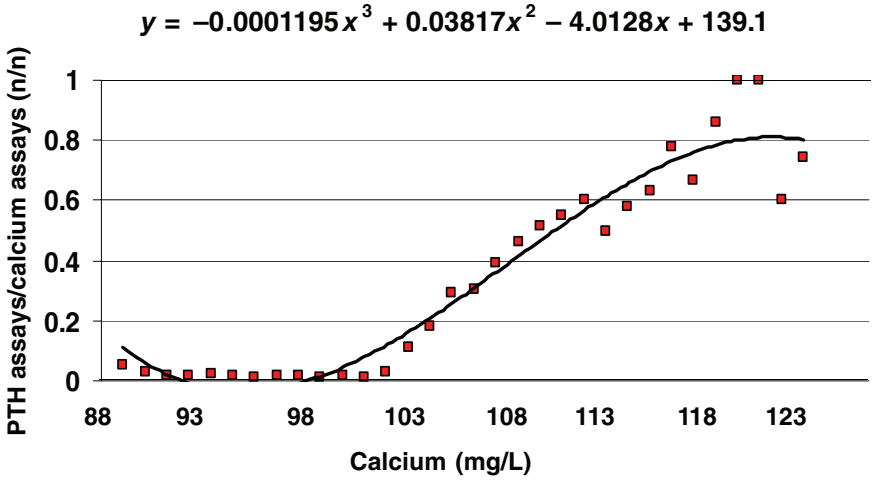

Fig. 2. Relationships between serum calcium concentrations and the frequency of follow-up requests for serum PTH measurements for female patients on Medicare.
Adapted from Gallaher et al. *(1)*.

These mathematical relationships between laboratory test values and the frequency of follow-up procedures do not provide explicit assay performance limits, but they provide a potential mechanism for analyzing the effects of analytic bias. If one assumes that the statistical relationships are at least in part causal relationships, then these curves of concentration vs frequency of orders can be used to simulate the effect of analytic bias. For example, in Fig. 2, if the assay for serum calcium is shifted upward by 2 mg/L, then patients with calcium values of 103 mg/L would have their calcium values reported as 105 mg/L. If the orders for serum PTH are triggered by serum calcium, 30% of these patients would have PTH tests, vs 20% based on the nonshifted calcium values. Integration over the full spectrum of calcium values would provide estimates of the impact of this shift in laboratory test results on the ordering of clinical procedures.

Approach 6: Analyte Performance Characteristics Based on Decision Models Used in Expert Systems

Analytic decision support systems have been developed to assist clinicians in patient care decisions. Two types of algorithms are commonly used in these decision systems: machine-learned algorithms based on data, and knowledge-engineered decision algorithms based on expert opinion and literature studies. The machine-learned algorithms generally have built-in uncertainty associated with the data, whereas the knowledge-engineered models often use fixed branch points and do not have uncertainty estimates unless they are explicitly incorporated *(34, 35)*.

A basic assumption in these decision models is that the analytic measurement systems are stable and not influenced by imprecision and bias, particularly at decision points. Bias increases false-positive and decreases false-negative results (or vice versa), whereas imprecision increases both *(36)*. Simulation studies have been used to analyze the effects of imprecision and bias on decision models. These studies add controlled error to the databases and analyze their impact on the decisions. These types of studies could be used to help define analytic performance goals *(37, 38)*.

Analysis of clinical-decision curves is a tool for evaluating some of the parameters of diagnostic tests *(39)*. This technique does not require explicit definition of external data costs, benefits, and preferences typically used in decision analysis techniques. Many traditional decision support systems use simple yes-no branch points with predefined decision points, such as hematocrit <30% or albumin greater than a median value *(40, 41)*. These types of models are less useful for assessing the impact of analytic performance.

Parametric statistical models can be developed for specific disease diagnoses by using the distributions of test values from patients with and without that specific disorder. Even simplified diagnostic classification systems using only 1 test and 3 states of health (hypofunction, normal function, and hyperfunction) require multiple data sets and various assumptions. However, when these models are developed they can provide a valuable framework for analyzing these effects of analytic bias and imprecision on medical decisions.

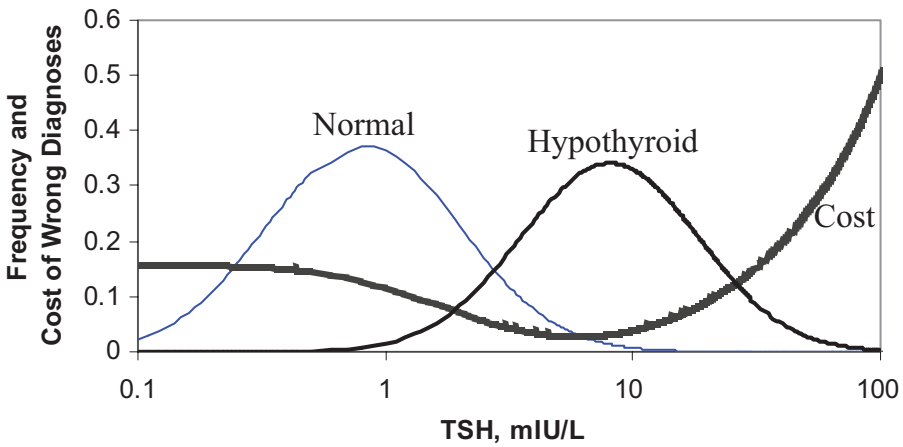

Fig. 3. Frequency distributions and cost optimization decision function for diagnosing hypothyroidism by using TSH measurements.

The nadir of the cost function occurs at TSH concentrations in the range between 5.0 and 10.0 mIU/L. These curves could be used to evaluate the effect of analytic performance on medical decisions. Reprinted, with permission, from Klee *(42)*.

The potential value of decision models for assessing analytic goals is illustrated with an example using TSH to classify patients into hypothyroid, normal thyroid, and hyperthyroid states *(42)*. The distributions of TSH values in the 3 disease states were developed along with estimates of the prevalence of hypothyroidism (5%) and hyperthyroidism (2%). It was assumed that the "cost" of false-positive diagnoses was equal for the 2 disease states and was assigned an arbitrary value of 10 units. The term "cost" relates to all the adverse consequences of these false-positive decisions, and the units are used only to provide a relative scale. It also was assumed that the cost for missing a diagnosis was dependent on how far the true TSH value was from the decision point. For hypothyroidism the cost of a false-negative result was assumed to be directly related to the TSH value, with a proportionality factor of 3 (cost = 3 × TSH value). Fig. 3 shows that the cost of misdiagnosis reaches a nadir for TSH values between 5.0 and 10.0 mIU/L. This relatively flat part of the curve could be used to assess the clinical impact of analytic uncertainty in TSH measurements in terms of the effects on diagnosis of hypothyroidism. For hyperthyroidism the cost of a false-negative result was considered to be inversely related to the TSH value (more cost for falsely low TSH values), and the cost curve showed a relatively flat nadir for TSH values between 0.1 and 0.2 mIU/L. These types of models potentially could be used to help define analytic performance limits that would be related to the medical utility of the tests.

Discussion

The goal of all clinical laboratorians is to provide the quality of results required for good patient care. However, quality is difficult to assess unless there are explicit performance criteria. Various methods have been proposed for establishing analytic performance criteria. The most widely used methods use biologic variation within patients and across groups of patients to provide statistical limits within which the analytic variation is masked by the larger biological variation. These limits work well for assessing assay imprecision, especially for medical decisions involving individual patients, but they do not address the performance issues that analytic bias may have on the aggregate system performance across large cohorts of patients. This review summarizes some alternative approaches to help set analytic performance goals in ways that are related to their impact on medical decisions.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. CDC. Current CLIA regulations (including all changes through 01/24/2004). http://www.cdc.gov/clia/regs/toc.aspx (Accessed November 2009).
2. Westgard JO, Klee GG. Quality management. 4th ed. St. Louis: Elsevier Saunders; 2006. p 485–529.
3. Fraser CG. General strategies to set quality specifications for reliability performance characteristics. Scand J Clin Lab Invest 1999;59:487–90.
4. Dhatt GS, Agarwal MM, Bishawi B, Gill J. Implementing the Stockholm Conference hierarchy of objective quality criteria in a routine laboratory. Clin Chem Lab Med 2007;45:549–52.
5. Fraser CG, Kallner A, Kenny D, Petersen PH. Introduction: strategies to set global quality specifications in laboratory medicine. Scand J Clin Lab Invest 1999;59:477–8.
6. Jones BA, Bekeris LG, Nakhleh RE, Walsh MK, Valenstein PN. Physician satisfaction with clinical laboratory services: a College of American Pathologists Q-probes study of 138 institutions. Arch Pathol Lab Med 2009;133:38–43.
7. Plebani M. Quality specifications: self pleasure for clinical laboratories or added value for patient management? Clin Chem Lab Med 2007;45:462–6.
8. Werner M. Linking analytic performance goals to medical outcome. Clin Chim Acta 1997;260:99–115.
9. Sciacovelli L, Secchiero S, Zardo L, Plebani M. External quality assessment schemes: need for recognised requirements. Clin Chim Acta 2001;309:183–99.
10. Westgard QC. Rilibak: quality goals the German way. http://www.westgard.com/rilibak-2.htm (Accessed March 2010).
11. Fraser CG, Hyltoft Petersen P, Libeer JC, Ricos C. Proposals for setting generally applicable quality goals solely based on biology. Ann Clin Biochem 1997;34(Pt 1):8–12.
12. Sciacovelli L, Zardo L, Secchiero S, Plebani M. Quality specifications in EQA schemes: from theory to practice. Clin Chim Acta 2004;346:87–97.
13. Gowans EM, Hyltoft Petersen P, Blaabjerg O, Horder M. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 1988;48:757–64.
14. Fraser CG, Petersen PH. Analytical performance characteristics should be judged against objective quality specifications. Clin Chem 1999;45:321–3.
15. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: pros, cons and progress. Scand J Clin Lab Invest 1999;59:491–500.
16. Barnett RN. Medical significance of laboratory results. Am J Clin Pathol 1968;50:671–6.
17. Skendzel LP. How physicians use laboratory tests. JAMA 1978;239:1077–80.
18. Skendzel LP, Barnett RN, Platt R. Medically useful criteria for analytic performance of laboratory tests. Am J Clin Pathol 1985;83:200–5.
19. Thue G, Sandberg S, Fugelli P. Clinical assessment of haemoglobin values by general practitioners related to analytical and biological variation. Scand J Clin Lab Invest 1991;51:453–9.
20. Skeie S, Thue G, Sandberg S. Patient-derived quality specifications for instruments used in self-monitoring of blood glucose. Clin Chem 2001;47:67–73.
21. Bais R. What information should manufacturers provide on their procedures? Clin Chem 2006;52:1624–5.
22. Burman KD. Commentary: discordant measurements of serum triiodothyronine (T4), thyroxine (T3) and thyroid-stimulating hormone (TSH). Clin Chem 2008;54:1246.
23. Snow CF. Laboratory diagnosis of vitamin B12 and folate deficiency: a guide for the primary care physician. Arch Intern Med 1999;159:1289–98.
24. Cohen J, Ward G, Prins J, Jones M, Venkatesh B. Variability of cortisol assays can confound the diagnosis of adrenal insufficiency in the critically ill population. Intensive Care Med 2006;32:1901–5.
25. Herold DA, Fitzgerald RL. Immunoassays for testosterone in women: better than a guess? Clin Chem 2003;49:1250–1.
26. Matsumoto AM, Bremner WJ. Serum testosterone assays: accuracy matters. J Clin Endocrinol Metab 2004;89:520–4.
27. Rosner W, Auchus RJ, Azziz R, Sluss PM, Raff H. Position statement: utility, limitations, and pitfalls in measuring testosterone: an Endocrine Society position statement. J Clin Endocrinol Metab 2007;92:405–13.
28. Vesper HW, Botelho JC, Shacklady C, Smith A, Myers GL. CDC project on standardizing steroid hormone measurements. Steroids 2008;73:1286–92.
29. Klee GG, Schryver PG, Kisabeth RM. Analytic bias specifications based on the analysis of effects on performance of medical guidelines. Scand J Clin Lab Invest 1999;59:509–12.
30. Shermock KM, Connor JT, Lavallee DC, Streiff MB. Clinical decision-making as the basis for assessing agreement between measures of the International Normalized Ratio. J Thromb Haemost 2009;7:87–93.
31. Klee G. A conceptual model for establishing tolerance limits for analytic bias and imprecision based on variations in population test distributions. Clin Chim Acta 1997;260:175–88.
32. Klee GG, Schryver, P. Quality Assurance for basic haematology cell counts. In: Rowan RM, van Assendelft OW, Preston FE, eds. Advanced laboratory methods in haematology, New York: Arnold; 2002. p 3–17.
33. Gallaher PM, Mobley RL, Klee GG, Schryver P, preparers. The impact of calibration error in medical decision making. Final report. [Gaithersburg, MD]: National Institute of Standards and Technology, Chemical Science and Technology Laboratory; 2004. Report nr: Planning report 04-1. Available at: www.nist.gov/director/prog-ofc/report04-1.pdf.
34. McNair P, Brender J, Talmon J. Computer-aided test selection and result validation: opportunities and pitfalls. Clin Chim Acta 1998;278:243–55.
35. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Med Decis Making 2008;28:650–67.
36. Petersen PH, de Verdier CH, Groth T, Fraser CG, Blaabjerg O, Horder M. The influence of analytical bias on diagnostic misclassifications. Clin Chim Acta 1997;260:189–206.
37. Egmont-Petersen M, Talmon JL, Hasman A. Robustness metrics for measuring the influence of additive noise on the performance of statistical classifiers. Int J Med Inform 1997;46:103–12.
38. McNair P, Brender J. Information enhancement in clinical decision making by controlled data generation. Scand J Clin Lab Invest Suppl 1990;202:112–9.
39. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53.
40. Chu A, Ahn H, Halwan B, Kalmin B, Artifon EL, Barkun A, et al. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artif Intell Med 2008;42:247–59.
41. Ritchie RF. A knowledge-based system to aid with the clinical interpretation of complex serum protein data. Clin Chem Lab Med 2001;39:1045–53.
42. Klee GG. Clinical interpretation of reference intervals and reference limits: a plea for assay harmonization. Clin Chem Lab Med 2004;42:752–7.