Adequacy of NCEP Recommendations for Total Cholesterol, Triglycerides, HDLC, and LDLC Measurements

To the Editor:
We read with concern the recent paper by Fallest-Strobl, Olafsdottir, Wiebe, and Westgard (1) in which the authors state, “... the NCEP recommendations fail to adequately consider the quality control requirements necessary to detect medically important systematic errors.” We also note that Westgard, Petersen, and Wiebe (2) published a similar paper in 1991 questioning the adequacy of the NCEP recommendations for total cholesterol. We appreciate the importance of the issues addressed in these two papers and find no fault in the majority of the derivations. We do notice, however, that the authors did not address the conditional nature of their calculations. That is, the OPSpecs charts do not take into account the fact that a patient result will only be reported when the quality control (QC) run associated with the patient specimen is considered to be in control. Thus, the OPSpecs charts relate to the conditional probability that a measured patient result will exceed a specified limit, not to the joint probability of patient misclassification. For misclassification to occur, the measured patient result must not only exceed the specified limit, but the QC procedure must also fail to indicate an out-of-control condition. Without taking into account the probability that the QC procedure will fail to detect an out-of-control condition, the authors do not accomplish their stated intention to evaluate the NCEP recommendations with regard to “quality control procedures that are necessary to detect unstable operation” (2). Incorporating QC performance into the OPSpecs charts requires inclusion of the probability that a patient result will be reported on the basis of the QC outcome of the analytic run in which the patient specimen was measured.

To illustrate our point, consider the intended 0.95 probability of correct classification chosen by the authors. This conditional probability does not reflect the overall likelihood of correct classification but rather applies only to the rare condition when the QC procedure fails to alert the analyst that the laboratory instrument may have been out of control during the run in which the patient specimen was analyzed. For instance, when evaluating systematic bias, the authors chose systematic errors detectable by QC with a probability equal to 0.90. The joint probability of correct classification in this case is actually 0.995 [1.0 – (0.05 × 0.10)]. That is, it is the complement of the product of the probabilities of two independent events: (a) The measured result will be beyond the critical decision limit (as defined by the authors, this event occurs with probability 0.05), and (b) the QC sample(s) measured during the run in which the patient specimen was analyzed will be within acceptable limits (on the basis of the example presented by the authors, this event occurs with probability 0.10). Similarly, when evaluating random error, the authors chose increases in random error detectable by QC with a probability equal to 0.80 (or 0.90). The joint probability of correct classification in this case is actually 0.99 (or 0.995). Thus, the conclusions reached by the authors about the adequacy of the NCEP recommendations are actually applicable to a more stringent correct classification requirement of at least 0.99 rather than 0.95 as they claim.

Addressing the subtle difference in meaning between the conditional probability inherent in the authors’ computations and the joint probability we recommend is somewhat difficult because the authors do not actually present their results in terms of computed probabilities. Instead, they present their results in terms of the maximum allowable inherent bias and/or analytic random error corresponding to a single fixed conditional probability of correct classification equal to 0.95. The corresponding joint probability associated with their results is actually much higher than 0.95. Both the conditional and joint probabilities are applicable to an individual patient who has a specified intraperson biological variation and whose health status is being determined on the basis of a single randomly collected specimen that is analyzed in a laboratory with a specified accuracy and precision and with a specified QC procedure. The joint probability provides an overall indication of the likelihood of correct patient classification given a specified systematic bias or increase in random analytic error. The conditional probability, on the other hand, provides an indication of the likelihood of correct patient classification only under the rare circumstance that the QC procedure fails to detect the specified systematic bias or increase in random error. In addition, the authors consider only one systematic bias and one increase in random analytic error: the minimum systematic bias that is detectable with probability 0.90 by any one of several QC procedures and the minimum increase in random analytic error that is detectable with probability 0.80 (or 0.90) by any one of several QC procedures. The conditional and joint probabilities could be evaluated using operating-characteristic curves to map performance characteristics as functions of inherent and systematic biases and inherent and increased random analytic errors in the context of various QC procedures and frequencies. The authors indirectly evaluated the conditional probability for one case of systematic error and one case of increased random error via their OPSpecs charts and concluded that the NCEP recommendations for precision and accuracy are inadequate.

We have developed operating-characteristic curves for LDLC, HDLC, total cholesterol, and triglycerides to evaluate the NCEP-recommended allowances for inherent bias and inherent analytic random error. These curves show that, for LDLC, HDLC, total cholesterol, and triglycerides, the current NCEP accuracy and precision recommendations are adequate to assure a high likelihood of correct patient classifications. Our analyses assume the use of standard QC procedures (e.g., Shewhart mean-and-range chart QC or Westgard multi-rule...
QC) that incorporate at least two QC pools with samples measured in duplicate. We suggest that at least two concentrations of QC material be included in the QC scheme to assure that the measurement system is operating within desired specifications across the entire range of normal and abnormal analyte concentrations and to assure with high probability that patients are correctly classified. We also suggest that intraperson biological variation can be reduced by obtaining two serial patient specimens at least one week apart. (3) The relative range of the two results can be used to determine whether additional patient specimens are required because of unusually high intraperson biological variation.

References

Samuel P. Caudill*
S. Jay Smith
Gerald R. Cooper
Gary L. Myers
Environmental Health Laboratory
Sciences Division
National Center for Environmental Health
National Centers for Disease Control and Prevention
4770 Buford Hwy. NE, F25
Atlanta, GA 30341-3724

*Author for correspondence. Fax (770) 488-4191; e-mail SPC1@CDC.GOV.

Two of the authors of the above-mentioned article reply:

To the Editor:
We agree that we are very demanding in our efforts to assure the quality of test results and to correctly classify patients. The concern of the authors (1) seems to be that we are more demanding than we say we are (0.99 vs 0.95). Our concern is that the NCEP specifications for method performance are not demanding enough to assure the quality required by CLIA proficiency testing criteria and NCEP’s own guidelines for clinical use of the test results (2–5). Although the NCEP specifications are only recommendations and not regulations, these recommendations still establish the target specifications of performance for analytical methods and systems, and if set too loose, they may inhibit initiatives to improve method performance.

We do not see the issue as one of unconditional vs joint probabilities. We want to make sure that specifications for method performance and quality control will assure that medically important analytical errors are detected by the QC procedures commonly used in laboratories today. One way of assessing the adequacy of the NCEP recommendations is to calculate the minimum or “critical-size” analytical errors that need to be detected on the basis of the quality required for the test and the imprecision and inaccuracy specified by NCEP. We consider errors of this critical size and larger to be medically important because they will cause >5% of the test results in an analytical run to exceed the defined allowable error. The probabilities of detecting these critical-size errors can be assessed from the operating characteristics or power curves for different QC procedures. When analytical methods just meet the NCEP precision and accuracy specifications, this assessment often shows that critical-size errors are difficult to detect by commonly used statistical QC procedures that have a low number of control measurements (such as the CLIA minimum of 2).

For example, given the CLIA proficiency testing criterion for acceptable performance of 10% for cholesterol and given a method that just meets the NCEP specifications of 3.0% imprecision and 3.0% inaccuracy, the size of the systematic error that needs to be detected equals 0.68 times the standard deviation (s) of the measurement procedure, i.e., a systematic shift of 0.68 × 3.0%, or 2.0%, will cause some test results to be in error by >10%. Will commonly used QC procedures be able to detect a systematic shift of 0.68s?

The power curves for commonly used QC procedures are shown in Fig. 1. The vertical line at 0.68 on the x-axis represents the critical-size systematic error, which is at the low end of the power curves. As can be seen from the intersections of the vertical line and the power curves, the probabilities for detecting a systematic error of this size are ~0.20 or less for commonly used control rules such as \(1.2\sigma, 1.5\sigma, 3\sigma, 2\sigma/R\), and \(1.5\sigma\) with 2 control measurements per run. The highest error detection is achieved with 2-s control limits, but false rejections are also high and will cause a waste of nearly 10% of the process output. The probability of error detection is only 0.10 or less for those QC procedures having acceptably low false rejections. In practice, if an accuracy problem occurs that causes a systematic shift just the size of this critical systematic error, there is only a ~10% chance of detecting the problem with commonly used QC procedures. Another way of expressing this outcome is that only 1 run in 10 that has errors this size will be rejected; the other 9 will be reported.

Of course, if the accuracy problem causes a much bigger error, say equivalent to 3–4 times the size of the standard deviation of the method, then the problem will most likely be detected, as shown by the probabilities at the high end of the power curves. From a design or planning point of view, it would be desirable if the imprecision and inaccuracy of a method were so good that a medically important error would always be a high multiple of the standard deviation of the method so it could be readily detected, i.e., the 2.0% cholesterol error needs to be a multiple of 3s, or for example, rather than 0.68s. The exact multiple de-
Depends on the particular QC procedure being used, e.g., there would be a 90% chance of detecting systematic errors of 2.6s, 3.0s, 3.2s, 3.7s, and $\sim 4.1s$ by the QC procedures represented in Fig. 1 ($1_{2x}, 1_{2.5x}, 1_{3x}/2_{3s}, R_{4s}, 1_{3w}$, and $1_{3.5w}$ respectively, all with $n = 2$). These values show the rather large “insensitivity” of commonly used QC procedures and the need to consider QC performance when setting method performance specifications.

In our approach to establishing operating specifications, we allow for these known insensitivities of QC procedures. We start by defining the quality required for the test, either as a total allowable error as defined by the CLIA proficiency testing criteria or as a medically important change or clinical decision interval based on the NCEP clinical interpretation guidelines. We then deduct preanalytical causes of variation (e.g., within-subject biological variation) when a clinical quality requirement is being used, take into account the size errors detectable with a probability of 0.90 or 90% chance of detection by different candidate QC procedures, and then calculate the imprecision and inaccuracy that would be allowable under stable operating conditions. The results of these calculations are presented in the form of an OPSpecs chart that displays the inaccuracy ($y$-axis) and imprecision ($x$-axis) that are allowable, given the quality required for the test and the known error detection capabilities of candidate control rules and numbers of control measurements.

One advantage of the OPSpecs tool is that the complex relationship between the quality requirement for the test, the imprecision and inaccuracy of the method, and the error detection of a QC procedure can be displayed in a very practical format. Given the observed imprecision and inaccuracy of a method, you only have to locate the “operating point” of that method (i.e., its imprecision as the $x$-value and its inaccuracy as the $y$-value) to identify appropriate QC procedures (by the control rules and ns corresponding to the lines that appear above the operating point). Given the QC procedures in common use, you can quickly assess the maximum allowable imprecision (for a bias of zero) from the $x$-intercepts of those lines; in addition, for any specified bias or $y$-value, the allowable imprecision can be determined from the corresponding $x$-value on the operating lines for the QC procedures.

We recognize that there may be alternative methodologies for developing specifications for method performance and quality control, but those methodologies are of limited value until they are published and made available to others. The approach used in our studies is well documented (2–5), readily performed with available tools and computer technology (6, 7), and supported by a web-based training course on “Quality Control Planning” that is available to anyone, any place, any time [http://www.aacc.org/westgard]. We think that our studies of the NCEP recommendations for lipid tests (2–5) provide evidence that laboratories need a practical methodology for establishing method performance and QC specifications.

References

Modern Quality Management Misunderstood?

To the Editor:

Currently, there is a trend in clinical chemistry to assess laboratory quality by so-called “quality management techniques”. These techniques enable managers to investigate the quality of complex processes and allow identification of weak points within these processes. In addition, they allow the investigation of patient-benefit-related outcome of testing. According to a recent editorial in this journal (1), the application of these techniques in the clinical laboratory are expected to yield “healthcare that is not only better but cheaper, and much more satisfying to practice.” I agree with that statement, but I was somewhat surprised that it was evoked by two studies in the same issue of the journal (2, 3) that, in my opinion, do not substantiate these expectations.

The studies that were cited, and nearly all studies of that kind, come down to the message that (a) the error frequency in the clinical laboratory is very low (2, 3); (b) most errors occur in the pre- and postanalytical phases (2); and (c) the vast majority of analytical errors would not have caused severe patient management problems (2, 3). In short, the reader is convinced that current analytical quality is excellent. On those grounds and considering the cost-pressure on the laboratory, nobody can take seriously such statements as “improvement (read: of analytical quality) should be possible” (3). I do not dispute the value of the cited studies, but their approach is limited. For analytical quality, they investi-

gated only whether the process had been applied correctly, and they assessed, in fact, relative quality. For example, in one of the cited articles, results were classified as unacceptable on the basis of the imprecision of the analytical methods used (3). Whether the imprecision itself was questionable was not discussed. In the other study (2), clinicians were asked to identify suspect results. I assume that the clinicians reacted on the basis of their experience with the analytical quality that was available. Whether the intrinsic quality of the test was satisfactory was not discussed. Thus, the weakness of such studies is that they do not recognize that error-free operation is meaningless when the intrinsic quality is poor. The same holds true when the clinical accuracy of a test is poor or when the test had been inappropriately ordered.

The cited articles, indeed, showed that process performance was excellent. However, the laboratory should also feel responsible for the intrinsic analytical quality it offers and for the value that a certain test has for the patient. I see danger in that the question of intrinsic analytical quality and test value will be pushed out of focus by such studies, and interest will be moved to the pre- and postanalytical phases. Indeed, the latter two might have been given too little attention in the past. Nevertheless, analytical quality should stay in focus, because it is the most important value the laboratory can offer. I feel it more urgent to locate the problem areas in the laboratory than to demonstrate that, in general, everything is perfect. Otherwise, old statements will come back in nice new clothes, such as the phrase: we were good, we are good, we will be even better in the future, and we only have to sell ourselves better.

Many of the problem areas are, in fact, known. Among them are measurements of free hormones or steroid hormones at low concentrations. In addition, many analytes are not unequivocally defined, and it is often not known what is really measured. Think, for example, of glycated polypeptides. Different tests give different answers, with the consequence that common reference intervals or cutoff values cannot be used. This will become a serious problem in the future because of the need for unified treatment strategies and the introduction of expert systems. For the same reason, standardization will become a major issue; in fact, it has not yet been achieved in many areas. Knowledge about internal quality control is still far from optimal and might even diminish in the future because of industry promises that the new systems have built-in quality control with no need for attention by the user. Reaction patterns when quality control rules are violated are often overly simplistic. For example, many people recommend remeasuring the control and, when it is “in again,” continuing with patient specimens.

Modern quality management, on the other hand, goes far beyond assessment of whether current processes are correctly performed. Its strength is its ability to disclose the weak parts of the overall process and to estimate the value of the process itself. However, this can be effective only when all input elements are checked for validity. In this view, modern quality management should assess actual quality on the basis of specifications for desired quality. Furthermore, it should provide tools that allow practitioners to anticipate future quality needs in an early stage.

Modern quality management is much more than the investigation of error rates and the effects thereof. The latter is valuable, but nowadays the more important problem for laboratories is to demonstrate that their services are useful for patient management. The primary task is not to prove that the measurements do no harm (which directly provokes concern that they are of no use either) but to demonstrate their benefits for the patient. Modern quality management should, therefore, refocus the laboratory on, for example, test selection. This needs another way of thinking, one that is primarily focused on the clinical utility of measurements. An exemplary article that demonstrates this kind of thinking.