Proficiency Testing as a Regulatory Device: A CAP Perspective
William Hamlin

The history and development of the proficiency testing programs of the College of American Pathologists are reviewed. Important considerations of external surveys include target value assignment and determination of acceptable ranges. “Blind” proficiency testing and on-site evaluation provide alternative methods of laboratory evaluation but are limited by practical logistics.

The College of American Pathologists (CAP) has been involved in interlaboratory comparison surveys since shortly after the first such survey was done in Pennsylvania in 1946 by Belk and Sunderman (1).

The first CAP subscription surveys were initiated in 1961 after multi-laboratory surveys in microbiology and in clinical chemistry had been distributed in 1959 and 1960, respectively. Since then, CAP surveys have grown to the point that 107 different surveys (24,359 subscriptions) are now shipped annually to laboratories in the United States, Canada, and 43 other countries. Coincident to this activity, the CAP has sponsored a series of national conferences on the subject (2–14).

The CAP has always perceived the survey programs as tools for interlaboratory comparison, and has consistently designated them as “surveys,” not as “proficiency testing.” The latter term did not come into widespread use until the 1970s, in conjunction with federal and state regulatory activities. This distinction, albeit semantic, is important when one considers current proposals to use interlaboratory-comparison data as the principal means of distinguishing “good” laboratories from “bad.”

The original and basic purpose of the College’s interlaboratory comparison survey programs was laboratory improvement through several more-specific objectives:

- To define the “state of the art” with respect to methodology, instrument/reagent systems, “accuracy” in reporting, etc. Thus, laboratories could be provided with information that would allow them to improve performance and to strive to meet the “standard of practice”
- To define sources of error in the analytical process and help to distinguish between systematic and random error
- To define criteria for substantiating or establishing medical usefulness parameters for reporting results
- To provide information that would facilitate achieving interconvertibility of laboratory data and results
- To assist laboratories in determining that their results may be unreliable in terms of the universe of laboratories, and thus to stimulate them to implement appropriate remedial and (or) educational measures that could improve performance

As of November 1991, neither the fundamental goals nor the more-specific objectives of the CAP surveys have changed since the inception of these programs. It has always been the position of the CAP that overall proficiency, or the lack thereof, is a consequence of multiple complex parameters—of which interlaboratory comparison is only one, and perhaps the least important.

In the mid-1960s the CAP Laboratory Accreditation Program instituted a requirement that a laboratory must subscribe to appropriate interlaboratory-comparison surveys as a condition for accreditation. This requirement was incorporated into the Quality-Control Standard as an “external” quality-control activity, not as a measure of “proficiency.” Those who devised the Accreditation Program were more interested in the remedial or corrective action taken by a laboratory that reported aberrant or deviant results on survey samples than whether it was always able to report “acceptable” results.

A laboratory’s true “proficiency” cannot be determined reliably through the use of proficiency testing (PT) or survey samples in and of themselves. PT samples are not the same as patients’ samples and cannot be handled in the laboratory in the same manner as patients’ samples. This reality is dramatically exemplified by the recently renewed and intense interest in “matrix effects” as they relate to analyses for cholesterol and other analytes in survey or PT samples. Advances in the technology for developing new instrumentation and instrument/reagent systems undoubtedly will magnify these kinds of problems.

The preparation and distribution of survey or PT samples to the diverse and widely dispersed population of laboratories in this country practically precludes the use of whole-blood samples or frozen components, at least for the present. As a consequence, PT sample preparation necessarily and unpredictably introduces variables not present in the usual laboratory sample (e.g., “matrix effect”).

Any consideration of the use of survey or PT samples as the principal determinant of laboratory proficiency as contemplated by the proposed CLIA ’88 regulations necessarily involves consideration of some critically important ancillary issues.

Value assignment. How is the target value to be determined? By analysis by the National Institute of Standards and Technology (NIST) to determine analytical “truth”? By referee or “Olympian” laboratories? If so, how many? Through the statistical determination of the all-participant, all-method mean? By using “peer group” means? Any one or all of these methods may be
appropriate in various circumstances; similarly, any one may be inappropriate in any particular situation.

The CAP has used referee or reference or "Olympian" facilities for determining target values. It still does so in certain disciplines; however, this methodology is not always adequate, as exemplified by the National Institute on Drug Abuse (NIDA) experience with its FUDT (Forensic Urine Drug-Testing) PT program. Target values were determined by a group of seven or eight referee laboratories, the high and low values were eliminated, and the mean of the remaining values was designated as the "true" value against which all participants in the program were measured. As it turns out, five or six of the referees were military facilities, all using the same methodology with the same instrumentation. This arrangement introduced a bias into the determination of the target value and many participants "failed," even though their results were perfectly valid. Because of the sensitivity of this particular area of testing, because of a number of threatened lawsuits, and because of an in-depth reanalysis, the method for establishing the target value has been changed and the all-participant mean is now used.

Grading. What is an "acceptable" or "passing" result? Is a reported result within two or three standard deviations around the mean indicative of "proficiency"? Is the use of "fixed criteria"—i.e., the target value plus or minus a percentage or a numerical value—the best way to evaluate results? Are "medical usefulness" criteria the preferable means of grading? If so, what are they and how does one determine them?

The CAP survey programs have tried all of these techniques at one time or another, with various degrees of success and failure. Again, any or all of these methods may be appropriate in various circumstances and within various disciplines. The College uses Resource Committees, each consisting of 6 to 12 "experts" in each laboratory discipline to design and evaluate every survey. They commonly alter the grading method for a given survey, depending on a variety of unpredicted circumstances or facts. It is not uncommon for the Committee to elect not to grade an analyte if the submitted results suggest that there has been a sample problem. In some situations a laboratory is using a particular instrument/reagent system that will not produce accurate results with a survey sample but performs reliably with patients' samples. In such cases, the use of peer group means is the valid way to evaluate those particular results.

Of more concern is that the grading mechanism, whether based on fixed criteria or standard deviations around the mean, may have no relevance to clinical situations or medical usefulness. The current regulations require that results for sodium must be reported within ±4 mmol/L around the target value. A laboratory that reports a result of 136 or 144 mmol/L on a sample whose target value is set at 140 mmol/L (in the mid-normal range) would "fail" and possibly would be required to cease testing for that analyte or to cease the subspecialty of clinical chemistry. From a clinical perspective, either result is irrelevant with regard to medical decision making with respect to electrolyte imbalance. The potential consequence of the laboratory being unable to measure and report results for sodium in a hospital environment is far more severe than ensuring that the results fall within the prescribed regulatory parameters.

Blind testing. Most would agree that the ideal mechanism for the periodic monitoring of a laboratory's performance would be a truly "blind" system. This would require the introduction of a sample that is indistinguishable from any other sample received by the laboratory. Such a sample would have to arrive in the laboratory from one or several of the sources that ordinarily submit samples to the laboratory, and would be routinely processed with other samples. The analyses would be performed and the results reported in exactly the same way as for any other sample. To establish such a system is not practicable, given the number of facilities that have to be dealt with and the magnitude of the distribution system necessary. Even more problematic is developing a mechanism for obtaining enough sample so that a valid target value can be defined, and ensuring that all participants receive a standard sample exposed to the same environmental and other variables so that the submitted results can be evaluated fairly.

Any compromise of the "ideal" erodes the usefulness of the "blind" system. Again, the NIDA experience is useful. They require the use of external blind samples as a condition of certification. Various providers of such samples, using various schemes for introducing the samples into the laboratory, have appeared, some of questionable reliability. However, the fact is that analysts in the FUDT laboratories can identify the "blind" samples with an alarming degree of accuracy.

The most reasonable compromise is an intralaboratory "blind" testing system, where actual patients' samples are reintroduced into the process without the knowledge of the laboratory personnel, reanalyzed, and reported. Many laboratories can and do successfully establish such internal systems, but they cannot be graded and they do not provide interlaboratory comparisons. Such systems do identify problems in the overall process, including the analytical process, and they do stimulate corrective measures—i.e., laboratory improvement.

On-site PT. One idea—that of having an "inspector" arrive in the laboratory, sample in hand, and watch the laboratory personnel perform an analysis to produce a result that is then used to determine proficiency—has long been advocated as a way to evaluate performance. There are several problems with this process. Certainly it distinctly deviates from testing "in the same manner as patients' samples are handled." Moreover, an entirely new set of variables is introduced, in addition to those inherent in mailed survey samples. The CAP tried on-site testing in conjunction with the Laboratory Accreditation Program in the 1960s and quickly abandoned the effort. The process is fraught with difficulties, including, but not limited to, maintaining the integrity...
of widely dispersed individual samples, establishing valid target values, maintaining valid chain-of-possession documentation, and having the inspector present and observing the process in widely differing testing schedules in each of the laboratories within the population of laboratories. There may be a potential role for this technique in situations where a laboratory has been identified as a problem facility by other means; however, nothing is gained through the use of on-site PT that cannot be gleaned more easily from other measures.

In summary, the College's basic position continues to be that interlaboratory comparison surveys are an important and valuable means for laboratory improvement. The information derived and disseminated by this tool is an invaluable component of laboratory quality-control and quality-assurance systems.

The CAP does not now, nor has it ever, considered that surveys or PT samples in and of themselves are appropriate means for determining whether or not a laboratory is "proficient." In our opinion, the tool is misused in the context of the proposed CLIA '88 regulations. The idea of rigid grading parameters with severe "sanctions," including prohibiting a laboratory from providing results in an analyte and/or subspecialty, is beyond the capability of, and contrary to, the reasons for doing interlaboratory comparisons.

Dr. Dennis Dorsey eloquently articulated our concerns in a paper he presented at the 1987 CAP Proficiency Testing Conference when he reiterated, "Using proficiency testing for law enforcement is like using a chisel to drive a screw. You can do it, but it doesn't work very well and it dulls the tool for jobs it can do better." I might add that the inevitable outcome is that the screws are poorly driven and you lose the use of the tool for its intended purpose. Is the identification of a very few "bad" laboratories, which most likely would have been identified by other means, worth the loss of a proven mechanism for wide-scale continuous laboratory improvement—namely, the interlaboratory comparison survey program?

References