Limitations of Proficiency Testing under CLIA '67

Ronald H. Laessig,1,3 Sharon S. Ehrmeyer,2 Barbara J. Lanphear,1 Barbara J. Burmeister,1 and David J. Hassemer1

Proficiency testing (PT), recognized as a quality-assurance (QA) and quality-improvement tool, has also become the cornerstone of the Health Care Financing Administration's (HCFA) regulatory strategy under the revised Clinical Laboratory Improvement Act of 1967 (CLIA '67) and the proposed Clinical Laboratory Improvement Amendments of 1988 (CLIA '88). Use of PT as a regulatory tool corrupts it for things it can do better. PT as a primary regulatory strategy has severe limitations. We explore the nature of these limitations and their implications for clinical laboratories as they impact on the long-term success of HCFA's approved regulatory PT programs in 1991 and beyond, and CLIA '88 PT, which is to be implemented in 1994.

Proficiency testing (PT) in U.S. clinical chemistry laboratories traces its origins to a fundamental study by Belk and Sunderman (1) in 1945. This often-cited work, as well as Sunderman's recent contributions (2), establish two concepts common to all PT programs: (a) initial experience often documents the existence of appallingly inferior (i.e., non-uniform) performance among participants, and (b) there follows, usually in the next shipment or "PT event," a remarkable improvement in uniformity of results. Most importantly, this improvement usually is sustained over the long term by the participants in the program. Today, PT programs could in substance—if not in overt design—be viewed as a prime example of a continuous, incremental improvement process (3). It is this realization that led the College of American Pathologists (CAP) to organize, promote, and in essence to mandate PT participation as a criterion for laboratory accreditation (4). During 1945–1967, PT participation became the common standard of practice in hospital and large reference laboratories.

In 1967, the U.S. Congress and the U.S. Public Health Service mandated, under CLIA '67, certain minimum performance standards for clinical laboratories engaged in interstate commerce (5). Subsequently, hospital laboratories receiving funding under Medicare were regulated along similar lines (6). The Centers for Disease Control (CDC), Atlanta, GA, as well as several individual states, provided interlaboratory, blind-sample PT programs. The CAP continued to provide its interlaboratory quality assurance (QA), eschewing the designation "PT," program.

Depending on local state regulations, during the period 1967–1991, the CAP program steadily garnered recognition as "acceptable for regulatory QA [i.e., PT] use" in the various states. The CAP never promoted its PT program for regulatory use, steadfastly claiming its purpose to be voluntary, educational, and directed at laboratory improvement (7–9). Until 1991, the CAP did not actually evaluate individual laboratory performance (at least for regulatory purposes), but did grudgingly agree to provide results, on request, to regulatory agencies to meet CLIA (or Medicare) requirements. Evidence of the success of the CAP program included the decision of the CDC to drop its parallel PT program in 1987. This decision caused many nonhospital reference laboratories and many state and local public health laboratories to look to CAP and (or) other state-run programs to fulfill the mandated PT requirement.

Limitation #1: When PT is used as a regulatory process to assess acceptability of a laboratory's performance, it loses its philosophical basis as a quality-improvement tool.

This regulatory recognition and widespread utilization did not come without a price. It leads to the first PT limitation. In a 1975 conference on PT, Dr. Dennis Dorsey, then CAP president, expressed the scientific community's concern when he said: "Using PT 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 the jobs it can do better" (10). At the same conference, Dr. Roger Gilbert, generally recognized as a major contributor to the philosophy behind, and operational logistics of, CAP's PT programs, said: "It is important that we do not allow any single perception of the application of interlaboratory testing to determine its use" (8). Dr. Joseph Boutwell, representing the CDC at the same conference (the CDC was the federal governmental body largely responsible for developing and implementing the regulatory strategy under the original CLIA '67 and Medicare rules) seemed to acknowledge the first limitation when he said, "At the time of the passage of CLIA '67, there was an assumption that proficiency testing was a fine and sharp tool for laboratory testing by which the quality of clinical laboratory performance could be monitored. It was also felt that such testing, coupled with licensure, would result in acceptable results in all interstate laboratories" (11).

This limitation came into clearer focus when, in 1989, Cembrowski and Vanderlinde (12) collected actual data
on the incidences of PT "cheating," that is, doing PT specimens in duplicate, among laboratories in Pennsylvania, and placed the rate at a minimum of 63%. Curiously, this "cheating" reflected nothing more than an intralaboratory attempt to optimize the quality (i.e., the accuracy) of its results. Belk, Sunderman, and their disciples would praise these individuals for seeking the truth (accuracy) by using the resources at hand and internally optimizing the value of the PT exercise. Today, HCFA suggests that a fine is the appropriate response (13). "Dulling" of the tool is, perhaps, an understatement.

Limitation #2: PT may not be the optimum tool for laboratory regulation circa 1991; unfortunately, it is the only tool.

Viewed in the perspective of history, mandatory PT under the original CLIA '67 and Medicare regulations (about 1968) more closely approximated a form of mandatory, external quality control (QC) than it did a rigorous regulatory process. The CAP's approach to regulatory agencies—"Here are the PT results: you grade them, we will not"—would seem to support the contention that the federal and state regulators as well as the CAP accreditation program, which had "deemed status," based self-initiated quality improvement on a combination of PT, QC data, procedural manuals, laboratory records, and most of all, on-site inspection and verification of acceptable intralaboratory practices. HCFA's CLIA '67 regulations (criterion E-32) actually reflected this: "The laboratory successfully participates in state-operated or state-approved proficiency testing programs..." (14). Because "successfully" was not explicitly defined in the HCFA inspector's handbook that interpreted the regulation, and since CAP (appropriately in our view) did not actually grade participants, few laboratories found themselves facing the cumbersome burden of a PT-induced regulatory action. Most laboratories used PT basically as a device to improve the quality of their performance.

The March 14, 1990, final rule under CLIA '67 dramatically changed this benign view of PT activities. This rule defined (a) explicit grading or performance criteria for 27 analytes in the subspecialty of routine chemistry (Table 1); (b) the PT format as five samples per shipment and four shipments per year, with a 45-day turnaround time for reporting results back to the laboratories; (c) acceptable performance requires 80% correct across all analytes in the subspecialty and four of five correct within each given analyte; (d) unsuccessful PT performance as failing the same analyte in two of three consecutive shipments; and (e) rigorous procedural requirements, i.e., prescribed and proscribed intralaboratory PT behaviors detailed below and in reference 15.

493.801(b) Standard; Testing of Proficiency Testing samples. The laboratory must examine or test, as applicable, the proficiency testing samples it receives from the proficiency testing program in the same manner as it tests patient specimens.

1. The samples must be examined or tested with the laboratory's regular patient workload by personnel who routinely perform the testing in the laboratory, using the laboratory's routine methods. The individual testing or examining the samples must attest to the routine integration of the samples into the patient workload using the laboratory's routine methods.

2. The laboratory may not test the samples with greater frequency of testing than it routinely tests patient samples.

3. A laboratory that performs tests on proficiency testing samples may not engage in any interlaboratory communications pertaining to the results of proficiency testing sample(s). Laboratories with multiple testing sites or separate locations may not engage in any communications or discussions across sites/locations concerning proficiency testing results.

4. The laboratory must not send the samples or portions of samples to another laboratory for analysis. Any laboratory that HHS determines intentionally referred its proficiency testing samples to another laboratory for analysis will have its approval and/or license revoked for at least one year. Any laboratory that receives proficiency testing samples from another laboratory for testing must notify HHS of the receipt of those samples.

5. The laboratory must document the handling, preparation, processing, examination, and each step in the testing and reporting of results for all proficiency testing samples and must maintain a copy of all records, including a copy of the proficiency testing program report forms used by the laboratory to record proficiency testing results, for a minimum of two years from the date of the proficiency testing event.

Perhaps the most onerous aspect of the CLIA '67 rule is the provision that if a laboratory fails in its analysis of one analyte—e.g., triglycerides—HCFA has the right to close down the entire 27-test subspecialty of routine clinical chemistry, which includes blood gases. Curiously, HCFA's public utterances on the topic during 1990 and early 1991 strongly indicated that they intended to do just that. At the August 1990, Congress on CLIA sponsored by the National Committee on Clinical Laboratory Standards, the "government" panelists gave no indication of a relaxation of this draconian policy (16). Paradoxically, on July 28, 1990, HCFA had approved the "sanction document" (published April 2, 1991), which espoused a much more lenient and thoughtful approach (13). We feel that HCFA finally had come to terms with the "Widespot, MN, Clinic" vs the "Mayo Clinic" phenomenon. Earlier thinking focused on the possibility that a PT failure would result in removing critical services from the geographic region of Widespot, MN, where the clinic is the sole provider of laboratory tests. Later it became apparent that the best large laboratories—e.g., the Mayo Clinic—characterized by excellent chemists using in-house-modified (optimized) equipment and methods, which focus on right answers—as opposed to "what everyone else gets"—were at greater risk in PT. Clearly, there is no backup system to replace the testing volume if a Mayo Clinic fails 4 of 10 triglyceride PT challenges in a
six-month period and must shut down all routine clinical chemistry testing.

In view of the above, and limitation #2, why pursue regulatory PT? HCFA readily conceded in the text associated with the CLIA '67 final rule that it was a prelude to the CLIA '88 rule. The CLIA '88 rule proposed in May 21, 1990, was virtually identical to the earlier offering in most operational areas, including PT. HCFA's intent was to put in place a mechanism to regulate not 12,000 hospital, reference, and some large clinic laboratories as under CLIA '67, but to deal with 100,000 to 600,000 new test sites, all of which fall under the mandate of the CLIA '88 amendments (16). We support the lower figure, based on reasonably accurate figures on actual numbers of test sites in Wisconsin. Even most conservatively, faced with a 10-fold (possibly 60-fold) increase in regulated test sites, HCFA cannot rely on regular on-site inspections and validation of internal QA practices to meet the congressional mandate to "assure consistent performance by [all] laboratories" (17). PT is the only viable option.

Limitation #3: PT standards have, de facto, determined minimum intralaboratory performance requirements for U.S. laboratories without considering the practical effect on these laboratories.

Our previous work (18, 19) has demonstrated that HCFA's minimum PT performance criteria can be translated directly into intralaboratory performance specifications, i.e., define the maximum allowable coefficient of variation (CV) for each analyte. In general, if a laboratory is able to achieve a day-to-day CV equal to one-third of the HCFA-specified PT criterion, it has virtually a 100% chance of passing the four of five within-analyte rule (19). More recently, we have considered the effect of HCFA's two overlapping PT requirements, i.e., four of five correct within an analyte to pass one PT event, and failure in a given analyte in two of three consecutive PT events, which constitutes grounds for HCFA to initiate an "adverse action" (20).

It should be noted, under CLIA '67, adverse action fully allows HCFA to order the complete cessation of testing in the entire subspecialty. Informally, HCFA has indicated they do not intend to take such an overt, onerous action except in extreme cases of fraud, abuse, or negligence. A step removed from entire subspecialty shutdown is "voluntary cessation of testing" in a particular analyte. This is definitely better; however, stopping blood pO2 evaluations certainly would adversely affect a hospital's ability to (e.g.) do surgery or operate an intensive care unit.

Under CLIA '67, which is fully in effect since January 1, 1991, and absent an official less rigorous regulatory strategy for sanctions as proposed for CLIA '88, HCFA has few alternatives except to decide to ignore PT data—which is precisely what it has done for all of 1991. Under the CLIA '88 proposal—specifically the April 2, 1991, sanction rule (18)—HCFA has indicated a much less draconian approach. A laboratory could file a "plan of correction," presumably self-imposed or HCFA mandated, to correct its PT deficiencies within a reasonable time. HCFA could also impose mandatory monetary penalties, from as little as $50 per day to a maximum of $10,000 per day, depending on the nature of the deficiency. HCFA would presumably cause the laboratory to be monitored at the laboratory's expense during the corrective process. Finally, HCFA reserves the right and decision to impose the "shut down" penalty.

Figure 1 shows, for a continuum of intralaboratory CVs, the probability that a laboratory will violate the four of five correct rule, i.e., have two or more incorrect results for one analyte in one PT event (rightmost, solid line), or for increasing numbers of multiple analytes, up to 27 (additional curves, right to left). The x-axis is the laboratory's internal CV; the CV is expressed as a function of HCFA's performance limits in Table 1. For example, HCFA's glucose PT limit is true value ±10%. Some laboratories have assumed that the 10% criterion implies HCFA is mandating that the laboratory's CV must be 10% or less. This is incorrect. As Figure 1 shows, for a single analyte, a 10% internal CV would result in a PT failure rate (two or more of five incorrect) of 51%, or failure of this analyte in approximately every other PT shipment. Figure 1 further indicates that if the laboratory's internal CV for glucose is reduced to 5%, or one-half of the HCFA PT tolerance limit, the chance of a single event failure is reduced to about 2%, or one in 50 PT events. Most importantly, if the CV is reduced to one-third of the HCFA limit, the chance of having one or more incorrect results in one PT event is virtually zero. Hence, our "rule of one-third" as a recommendation for an internal performance target to pass PT—provided a laboratory's bias is zero (19).

Still focusing on Figure 1, if a laboratory is testing two analytes, both of which have a CV equal to one-half the HCFA limits—e.g., a CV for both glucose and cholesterol of 5%—the second curve from the right shows that the risk of a single PT event failure increases from 2% to 4%. Three conclusions related to an individual laboratory follow from these curves: (a) reducing all CVs to less than one-third of the HCFA limits results in
essentially a zero risk of PT event failure, (b) having more than one analyte with a CV greater than one-third of the HCFA limit causes the failure rate to increase, cumulatively but not linearly, and (c) the worst analyte(s) in terms of the relationship of its CV to the HCFA limit will largely determine the PT failure rate (19). For example, if 26 analytes meet the one-third or less requirement, and the 27th does not, the rightmost curve for 1 analyte will predict the laboratory's PT performance, or failure rate. If results for 25 analytes are good but 2 are not, then the curve for 2 analytes can be used to predict the failure rate.

Figure 2 incorporates a coexisting bias of 20% of the HCFA PT limits while assuming the same CV relationship as discussed for Figure 1. For glucose with a 10% limit (Table 1), a 20% bias represents 0.11 mmol/L (2 mg/dL) when the target value is 5.55 mmol/L (100 mg/dL). Note that the starting point of the curves shifts from 33% to 28%. In Figure 3 the bias is 50% (i.e., 0.28 mmol/L (5 mg/dL) for the glucose example) of the HCFA limit; the starting points shift to the 20% point on the x-axis. This illustrates that coexisting bias reduces the "allowable" CV from 33% to 20% of the HCFA PT limit. For glucose, a 50% bias reduces the allowable CV to 2%, an impossibility for most laboratories. See the discussion related to Tables 3 and 4 for practical examples.

We suggest that laboratories look at the values in Table 1, divide each by three, and compare the results with their routine day-to-day CVs. For those tests where internal CVs exceed one-third of the HCFA limits, troubleshooting or possibly method changes are in order.

Limitation #4: PT specimens introduce idiosyncratic bias into the regulatory process.

We use the term "idiosyncratic bias" to represent that bias (with respect to the reference method) introduced into the analytical results by specimens whose matrix varies significantly from that for which the analyzer was designed.

The PT process, as mandated by HCFA in CLIA ’67 and CLIA ’88, brings into absolutely clear focus two irreconcilable conflicts. Manufacturers of today's instruments and reagent systems design and then manage and control the manufacturing process to ensure consistent results on fresh, human specimens. On the other hand, manufacturers do not design systems for the purpose of analyzing "control sera," which is essentially a euphemism for PT specimens. In general, these PT sera often are non-human-based (bovine) and almost always contain extracts of many species—e.g., chicken-intestine alkaline phosphatase, etc.—to augment endogenous concentrations of analytes; typically have been frozen and thawed multiple times in the manufacturing
process; and finally have been dispensed and lyophilized under less than ideally uniform conditions. They are then reconstituted in the laboratory with the use of various water sources, temperatures, volumes, and degree of agitation and aeration. The resulting variation in performance across specific methods is recognized in similar products we define as "controls with assigned values" with labeling that reflects many different "target values" for each analyte. Table 2 is such a listing of values for a well-understood analyte, chloride, on one of these products. Manufacturers of instruments address this problem by "assigning" analyte values to these products by a four-step process: (a) analyze a group of fresh, human specimens by the Reference Method; (b) tentatively set the instrument's calibration by use of reference values or reference specimens; (c) analyze the same group of fresh human specimens; and (d) adjust the instrument's calibration values and repeat the analysis until the system, using the "assigned" value, yields the Reference Method result for fresh human specimens. These assigned values often vary from system to system (Table 2), whereas both systems claim, and achieve, correspondence to the Reference Method for fresh human specimens.

These methodological differences typically are handled in PT programs by grouping similar methods for grading purposes. Unfortunately, HCFA has been less than totally lucid on the issue of how PT program providers may group similar or like methods for grading purposes. In the Wisconsin PT program that HCFA approved for 1991 under CLIA '87, we vigorously attempted to determine appropriate subgroups for PT grading purposes (21). In the cited reference, we demonstrated that in a high-quality laboratory, a well-characterized, commercial cholesterol method known to be correctly calibrated to yield results invariant from the Abell-Kendall Reference Method (on fresh, human specimens) repeatedly failed PT when graded either against the Reference Method or the "all enzymatic cholesterol methods" group mean. Table 3 shows PT performance data from seven samples from recent CAP surveys. For six of the seven PT samples, the Kodak Ektachem 500 gave results higher than the group mean. The Ektachem 500 instruments constituted ~2% of the total laboratory population, meaning that it did not greatly influence the group mean value. With respect to the Reference Method, the Ektachem 500 mean value was the same for one PT specimen, higher for four, and lower for two. In three of the cases, the results were >5% higher, i.e., representing an idiosyncratic bias of 5% (or 50% of the HCFA target tolerance), which automatically reduces the allowable internal CV consistent with passing PT to less than 2%. An internal CV of this magnitude is hardly achievable, making it nearly impossible for laboratories using this instrument to pass PT consistently. Yet, these laboratories have no alternative but to risk flunking PT, because to attempt to "adjust calibration points" to try to find the "all enzymatic methods group mean" for the PT material is shooting in the dark.

A similar situation exists for glucose testing in PT programs. Table 4 shows data from one specimen in the 1991 CAP survey. Again, two of the three competent methods, all calibrated to the Reference Method, are in serious trouble with respect to PT performance. One system has a bias of 5% and one 3%. Both biases significantly reduce to impossibly small values the respondents' allowable intralaboratory CV required to consistently pass PT. On the basis of limitation #4, and realizing that PT specimens in most programs, including Wisconsin's, will continue to be commercial control sera, we consider it mandatory to rigorously seek to appropriately, and optimally, group methods for grading purposes.

Limitation #5: HCFA's performance limits, while empirically determined, do not systematically relate to the current state of the art in real laboratories.

CAP's survey programs, from that of Sunderman in 1947 and 1991 (1, 2) to that of Hamlin in 1991 (4) have enunciated national goals for PT. The first-mentioned goal is "to define the state-of-the-art with respect to methodology, instrument/reagent systems ..." Though
Table 5. Proficiency Testing: CV vs HCFA Limits

<table>
<thead>
<tr>
<th>Test</th>
<th>Bias = 0%</th>
<th>Bias = 20%</th>
<th>Bias = 50%</th>
<th>Kodak acs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>4.0</td>
<td>3.3</td>
<td>2.0</td>
<td>5.5</td>
</tr>
<tr>
<td>Total bilirubin</td>
<td>6.0</td>
<td>6.7</td>
<td>4.0</td>
<td>8.8</td>
</tr>
<tr>
<td>Creatinine</td>
<td>4.0</td>
<td>5.0</td>
<td>3.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Ca</td>
<td>4.0</td>
<td>3.3</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Glucose</td>
<td>4.0</td>
<td>3.3</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Mg</td>
<td>10.0</td>
<td>8.3</td>
<td>5.0</td>
<td>5.4</td>
</tr>
<tr>
<td>K</td>
<td>2.7</td>
<td>3.3</td>
<td>1.4</td>
<td>2.3</td>
</tr>
<tr>
<td>TP</td>
<td>4.0</td>
<td>3.3</td>
<td>2.0</td>
<td>2.7</td>
</tr>
<tr>
<td>Na</td>
<td>1.1</td>
<td>0.9</td>
<td>0.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Urea N</td>
<td>3.6</td>
<td>3.0</td>
<td>1.8</td>
<td>4.2</td>
</tr>
<tr>
<td>Uric acid</td>
<td>6.8</td>
<td>5.7</td>
<td>3.4</td>
<td>7.0</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>4.0</td>
<td>3.3</td>
<td>2.0</td>
<td>3.1</td>
</tr>
<tr>
<td>HDL*</td>
<td>9.6</td>
<td>8.0</td>
<td>4.8</td>
<td>7.1</td>
</tr>
<tr>
<td>TG†</td>
<td>7.2</td>
<td>6.0</td>
<td>3.6</td>
<td>4.0</td>
</tr>
</tbody>
</table>

* These data from second CAP survey of 1991.
* HDL based on an assumed grading group CV or 8.0%.
* TG based on an assumed grading group CV of 6.0%.

Abbreviations as in Table 1.

1242 CLINICAL CHEMISTRY, Vol. 38, No. 7, 1992

We recently studied the effect of HCFA's PT rules on long-term or steady-state performance. In essence, this work (20) determines by direct calculation the probability of failing, first, the four-of-five correct rule, as a function of analyte intralaboratory CV over an infinite series of single PT events and then, subsequently, the probability of failing the two-of-three-consecutive-event rule over the same infinite series. The results are shown in Figure 4 (a–c), which can be interpreted in two ways. The y-axis is the percentage of laboratories in one of three PT status categories—successful, on probation (one PT failure for at least one analyte in the current or previous PT event), or in suspension status (at least one two-of-three-consecutive-events failure). Alternatively, the y-axis is the probability that a laboratory will find itself in a particular performance status once it reaches steady state, i.e., participates in at least eight PT events. The x-axis is the same as Figures 1–3: the intralaboratory CV expressed as a percentage of the HCFA PT performance limits in Table 1. Curves a, b, and c represent, respectively, coexisting bias values of 0%, 20%, and 50% of the values in the HCFA table. Only the curves for a single analyte are shown; these same curves reflect the performance of the worst analyte, the one that determines the laboratory's PT fate (20).

By focusing on curve 4a, it is readily apparent that the "no risk" point—i.e., the point at which the "probation" curve rises above the zero point on the y-axis—is at 40%, not the previous 33%. This is due to the concurrent application of the four-of-five-correct-results requirement for each analyte and the two-of-three-consecutive-events rule. It is for this reason that we used 40% in Table 5 for the zero bias case. Further, by observing that the "suspended" curve does not rise above zero until ~50%, we can gain some insight into the effect of HCFA's dual requirements. We did not complete our study or publish our technique until after publication of HCFA's PT performance criteria. We assume that HCFA has not undertaken such an analysis, prospectively or retrospectively, and their PT limits appear to have been against an absolutely homogeneous group, e.g., Ektachem against only Ektachem; acs against only acs groups. Despite this, for both instruments the group SDs exceed the "40% of the HCFA criteria" for 5 of 14 analytes. In the context of Figure 1, at least five analytes have CVs to the right of the critical point; hence, there exists a finite probability of PT failure, which increases dramatically with the number of analytes. If the group SD did in fact represent a typical laboratory under normal operating conditions, these laboratories—absent invoking some HCFA-precluded protocol to reduce imprecision such as doing PT in duplicate—are playing Russian roulette, with 5 of the 14 cylinders containing live ammunition.
selected empirically. Our study's conclusions can be summarized as follows. Assuming zero bias, if a laboratory reduces its internal CVs to 40% of the HCFA-table values, it will essentially avoid probation. Assuming CVs of 50% or less of the HCFA-table values, the laboratory may find itself on probation from time to time but rarely, if ever, in suspended status. As discussed (19, 20), multiple analytes that exceed the critical 33% or 40% values increase the overall jeopardy of the laboratory. One could conclude that HCFA intended for laboratories, depending on whether the possibility of being in probationary status is acceptable, to achieve internal performance standards represented by 4% or 5% CVs for glucose and cholesterol, 1.6 or 2.0 mmol/L for sodium, 2.05 or 2.56 μmol/L (0.12 or 0.15 mg/dL) for bilirubin, etc. In essence, these are de facto national performance standards. In all cases, the 40% figure is probably the operative one; few laboratories would enjoy being on probation, with a threat of suspension hanging over their heads.

Figure 4b and c, as well as the "bias = 20%" and "bias = 50%" columns in Table 5, show the effects of coexisting bias. As discussed under Limitation #4, for PT purposes an idiosyncratic bias and true bias (i.e., systematic error) have the same net effect on the grading of results. At 20% bias, the Ektachem still has 5 analytes out of 14 at risk; the ocu has 6 of 14. Previously, we concluded that a bias of 20% or less presented little PT risk; the data in Table 5 support this to some degree. A bias of 50%, on the other hand, presents major difficulty. Both instrument groups, at 50% bias, are at risk on 14 of 14 analytes studied. If the "bias = 50%" column is read as maximum allowable intralaboratory CV, it is readily apparent that such CVs are too small to achieve. Extending the CAP argument that the group CV represents the state of the art, the presence of 50% bias seems to imply that, to pass, a laboratory should be twice as good (i.e., have one-half the imprecision) as its peers. Refer again to Table 4; when graded against the "all automated chemistry" mean, one system starts with 50% of the HCFA limit bias; one with 30%.

Conclusions

The use of PT, or interlaboratory QA surveys for regulatory purposes obviously perverts their original intended use. Further, it precludes laboratories—by defining heretofore typical PT-related practices designed to improve performance (and hence accuracy) as "cheating," and subject to fines—from maximizing the benefits of the effort expended to improve performance. The draconian consequences of failure, i.e., potential suspension of testing in an entire subspeciality or even one analyte—a right HCFA has chosen not to renounce—hangs over the head of laboratories like the sword of Damocles. In almost all instances, if a laboratory tries to learn something about a particular instrument or method through PT, it could be accused of a CLIA infraction. If it pursues the learning opportunity, it had better be able to readily prove that only a single pass through a routinely used instrument took place and that the results were impervious to post-analysis influence. Before pursuing the "educational" opportunity associated with a PT event, PT data should be signed, sealed, and mailed out of the laboratory. This may be more bother than it is worth. On the other hand, this is also not what happens, routinely, to typical patients' specimens—but that's another story. All of this borders on nonsense—a nonsense born out of displacement of the fine QA tool from its intended purpose. Basically, a host of problems are going to continue to plague the regulatory PT as it is forced to meet performance expectations that it is philosophically, intellectually, and practically not capable of providing.

Voluntary, and even regulatory, PT has a strong potential to move the laboratory community toward a
fundamental accuracy base. The presence of idiosyncratic bias does not preclude this as an objective. If the laboratory community realizes that integrated instrument/reagent/calibrator and control systems—i.e., "closed systems"—are allowed in PT and compared with target method-specific values (in essence, homogeneous group means), the use of PT to achieve a true concept of accuracy is still viable. Logically, the manufacturer designs a system, the accuracy of which is based on reference methods, that achieves quality results on the "designed-for" matrix, i.e., fresh, human specimens. By grading PT results against the highly homogeneous group means, the limitation of PT specimens, specifically idiosyncratic bias, is largely overcome. In essence, by allowing idiosyncratic bias to be included in the target value, the laboratory is not penalized by the idiosyncratic behavior induced by the PT specimens. As a matter of practice, manufacturers are predisposed to cross-correlate their lots of calibrators, with use of fresh human specimens, to the reference methods. A calibration system, even one with specific (different) calibrator values for different lots of reagents, packs or slides, etc., so used, is "anchored" and traceable to a fundamental accuracy point. Conversely, in a situation where a manufacturer attempts to trace accuracy through a commercial control pool or, worse yet, through PT target values derived from reference methods, the manufacturer and the laboratory run a high degree of risk in HCFA PT programs.

The specific group mean is the logical choice for use in grading at this stage of PT evolution, but there are better ways. Use of a carefully selected homogeneous group of instruments with controlled lot numbers of calibrators and reagents could better serve as a referee laboratory system to determine target values. The referee group mean (target value) could be managed to provide better traceability to the manufacturer's anchor system, i.e., the Reference Method. This modification of grading is, to those amenable to a broad interpretation of the HCFA rules for PT providers, a viable possibility. Establishing and managing such a group of referee laboratories will take a rethinking of the common wisdom that manufacturers and regulators are natural enemies: the manufacturer must be allowed to participate, even coordinate the process.

Finally, we alluded to the fact that it appeared that the PT limits were not derived on a theoretical basis but rather on an empirical approach. This does not address the underlying fundamental issue of performance limits based on medical need or medical usefulness. Our German colleagues, primarily Professor D. Stamm, inaugurated such an approach with their regulatory PT programs (23). They do not, generally, appear to deal adequately with the issue of idiosyncratic bias, preferring to grade on the basis of Reference Method values assigned to PT specimens—which, in our opinion, still exhibit both method-related differences and idiosyncratic bias. There appears to be an opportunity for both sides of the Atlantic to learn.

References
16. Executive Summary, National Congress on CLIA-88, hosted by the National Committee for Clinical Laboratory Standards, August 6–8, 1990, Arlington, VA.