The Validity of Proficiency Testing: A Review of Data from the College of American Pathologists Laboratory Improvement Programs

Noel S. Lawson, M.D.*
Department of Pathology
St. John Hospital and Medical Center and
Wayne State University School of Medicine
Detroit, Michigan

Dan Tholen, M.S.
Statistical Consultant
Traverse City, Michigan

John W. Ross, M.D.
Department of Pathology
Kennestone Hospital
Marietta, Georgia

George G. Klee, M.D.
Department of Pathology and Laboratory Medicine
Mayo Clinic and Mayo Foundation
Rochester, Minnesota

*** Presenting Author**

Abstract: The validity of proficiency testing (PT) can be studied through comparing and correlating participant performance with that measured in other laboratory evaluation programs, and by examining consistency of grading. The College of American Pathologists (CAP) has undertaken various studies assessing the CAP Surveys PT Program. Performance correlates significantly with participation in the CAP Laboratory Accreditation Program (LAP), with LAP participants achieving overall lower rates of unacceptable results. For a set of analytes studied there is moderate positive correlation between bias and precision performance as measured by PT compared to that determined through CAP Quality Assurance Service (QAS) affiliated regional internal quality control. Furthermore, QAS participants achieve better survey performance than nonparticipants. Significant correlation has been demonstrated between performance measured in the CAP Linearity Survey and by concurrent PT. Relatively high consistency of participant survey performance ranks over time has been documented for two three year testing cycles. Finally, continuous improvement over several years has been documented for laboratories participating in the EXCEL Survey, with more experienced laboratories achieving significantly lower rates of unacceptable results. The findings, in aggregate, support the validity of PT, in the context of multiprogram characterization of laboratory performance.

Introduction - Validity of PT

Through the Surveys Program, the College of American Pathologists (CAP) is a dominant provider of professionally directed clinical laboratory Proficiency Testing (PT). This presentation focuses on contributions by CAP assessing the value of PT, emphasizing significant interprogram qualitative and quantitative relationships between PT and alternate measures of laboratory quality, consistency of performance, and effects of experience and time on results. Both previously published and new data support the validity of PT.

Material and Methods

In published studies, Proficiency Testing (PT) performance results have been compared with data from regional internal quality control, laboratory accreditation,[1] and linearity/calibration studies.[2] Tholen et al have reported the effect of experience and length of participation on PT performance.[3] We now provide additional data on the relationship between performance in PT and participation in the CAP laboratory accreditation program (LAP) and on consistency of performance in PT over time. To compare PT performance with LAP participation, survey performance from 1988-1990 was quantitated using an index which reflected aggregate score for 10 chemistry analytes over 3 years. Two sets of five common analytes each were used. (Set 1 = Calcium, Cholesterol, Creatine Kinase, Glucose, Potassium; Set 2 = Aspartate aminotransferase [AST], Bilirubin, Creatinine, Sodium, Triglycerides). Index scores P11, P12, P21, and P22 were derived from a nonparametric algorithm designed to penalize large bias and poor precision, respectively, and ignore small deviations from the target. All were derived from the same formula, with P12 and P22 giving slightly higher penalty for poor precision. P11 and P12 were for the 1st set of analytes and P21 and P22 were for the second. Index scores were prepared for five categories of laboratories, i.e., hospitals <100, 100-500, and >500 beds, independents, and others. For a laboratory to be included as an LAP participant, it must have been enrolled in the program for at least 1 year during the 1988-1990 period. In addition, a study of performance covering 1991-1994 was performed. All laboratories that were inspected (on-site) and active before 1989-1994 were included. A laboratory was considered to be accredited, i.e., "LAP" for the year of inspection and the following 2 years. Two quantitative survey performance measures were used - i.e., the Rate of Unacceptable Results and the Average Percent of Allowable Deviation (PAD), both using limits established by the Health Care Financing Administration (HCFA).[4] Unacceptable rates were obtained for chemistry, bacteriology, and immunology (qualitative and quantitative) challenges, which included all commonly performed HCFA regulated analytes.[4] The rate was obtained for all graded specimens in the specialty for each of the four years. Average PAD was determined for quantitative analytes in chemistry and immunology. Differences in means were analyzed by the Wilcoxon two sample test.

To study consistency of performance, Hematology (H1,H2) and Immunology (S,SM) Surveys data for 16 analytes from 1988-1990 were analyzed for laboratory performance. The study included 2736 participants with at least 20 challenges per survey per year, and at least 100 challenges overall. Performance cutpoints were set at the 25th and 75th percentiles of unacceptable

LAP LABORATORIES vs SURVEY LABORATORIES WITH AT LEAST ONE UNACCEPTABLE RESULT*

	No. of Laboratories				
ANALYTE	TOTAL	NOT IN LAP	WITH UNACCEPTABLE RESULTS,(%)	IN LAP	WITH UNACCEPTABLE RESULTS,(%)
AST	5071	2838	850 (30.0)	2233	606 (27.1)†
GLUCOSE	7718	4675	502 (10.7)	3043	177 (5.8)‡
PHOSPHORUS	4555	2009	305 (15.2)	2546	215 (8.4)‡
POTASSIUM	7459	4443	249 (5.6)	3016	92 (3.1)‡

*LAP indicates Laboratory Accreditation Program; AST, aspartate aminotransferase
†P<.05 by x^2 ‡P<.0005 by x^2
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 1

rate. The 3 years' consecutive performance was ranked 1 = top quartile, 2 = middle two quartiles, and 3 = lowest quartile. The 27 possible combinations were analyzed for actual vs. predicted performance. The study compared the observed rates of participants who were consistently in the same quartile group, with rates that would be expected if performance class were random.

The 1988-90 study was recently updated using data from Chemistry, Hematology, Immunology, and Bacteriology Surveys. These were analyzed to evaluate consistency of performance from 1992 to 1994. Participants were divided into three groups according to performance, i.e., relatively high = zero unacceptable results, relatively low = approximately 0-20 percentile, and intermediate = all others. The expected performance over 3 years was likewise the product of the percentages of participants in that group in each of the 3 years.

Results and Discussion

Using data from 1986 CAP programs, Lawson et al compared performance in PT by laboratories in the CAP LAP vs. that of nonparticipants.[1] They examined the analytes aspartate aminotransferase (AST), glucose, phosphorus, and potassium. The point of separation was one or more unacceptable results for an analyte during the year. In all cases, significantly more laboratories with unacceptable results joined the group of LAP nonparticipants (Table 1). The 1988-1990 data, using aforementioned indices, confirm that LAP participation is associated with improved PT performance. This is manifest as lower scores on multiple indices, across all categories of laboratories. Table 2 summarizes data on the relative performance of LAP and non LAP Survey participants. For all categories of laboratories as well in aggregate, LAP participants have better survey performance. The differences in performance are most

1990 SURVEYS PERFORMANCE INDICES VS. LAP STATUS
FOR EACH INSTITUTION TYPE AND ALL LABORATORIES†
(Lower index indicates better performance)

PRIVATE, COMMUNITY & FEDERAL HOSPITALS

PERFORMANCE MEASURE	1-99 BEDS NOT IN LAP	1-99 BEDS IN LAP	100-500 BEDS NOT IN LAP		100-500 BEDS IN LAP	500+ BEDS NOT IN LAP		500+ BEDS IN LAP
	n=932	n=288	n=787		n=1486	n=83		n=300
Index P11	73.1	70.6	61.3	*	55.9	61.2	*	56.5
Index P12	79.3	76.8	66.1	**	60.1	65.2		60.9
Index P21	69.0	67.9	58.3	**	54.8	70.2	**	55.3
Index P22	74.8	73.7	63.1	**	59.3	74.8	**	59.8

	INDEPENDENT NOT IN LAP	INDEPENDENT IN LAP	OTHER NOT IN LAP		OTHER IN LAP	ALL LABS NOT IN LAP		ALL LABS IN LAP
	n=243	n=146	n=251		n=52	n=2552		n=2402
Index P11	69.6	** 55.3	68.0	*	61.1	68.2	**	58.1
Index P12	75.3	** 59.5	73.6		67.1	73.7	**	62.7
Index P21	71.0	** 62.7	68.4		64.2	66.1	**	57.2
Index P22	76.9	** 67.5	73.9		69.5	71.5	**	61.9

STATISTICAL SIGNIFICANCE NOTES: * = .01 < P < .10
(Wilcoxon 2-sample test) ** = P < .01

†LAP = Laboratory Accreditation Program

Table 2

significant in medium-sized and large hospitals, and independents, and in the all laboratories grouping. Of the various groups studied, the lowest indices were noted among the medium-sized and large hospital cohorts.

The third surveys vs. LAP study, using the 1991-1994 data, has reaffirmed the better PT performance of LAP participants. In all specialties and years, LAP accredited laboratories have significantly better survey performance than non-LAP laboratories (Table 3a-d). This difference is seen both in the data reflecting rates of unacceptable results as well as in the PAD.

Thus, three different studies, each using different survey performance endpoints, yield the same conclusions. Laboratories in the CAP Surveys who are also in the CAP LAP program obtain better performance than non-participants. These studies have not been designed to evaluate the sources of improved performance. Possible contributing factors include directorship, overall attention to quality and documentation, adherence to the specific quality-related LAP questionnaire

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 CHEMISTRY

YEAR	NON LAP			LAP		
	NO.	UNACCEPTABLE (%)	ERROR	NO.	UNACCEPTABLE (%)	ERROR
1991	3358	2.62	40.0	2390	1.36	34.3
1992	3417	2.19	38.1	2597	1.08	32.2
1993	3453	1.89	36.1	2843	0.80	30.0
1994	3139	1.71	34.8	2906	0.74	29.2

All mean differences significant $p < .01$

Table 3a

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 BACTERIOLOGY

YEAR	NON LAP		LAP	
	NO.	UNACCEPTABLE (%)	NO.	UNACCEPTABLE (%)
1991	3012	7.60	2562	4.37
1992	2645	6.88	2571	3.74
1993	2482	6.52	2644	3.90
1994	2106	4.52	2679	2.94

All mean differences significant $p < .01$

Table 3b

items, and the integrated requirement within LAP that PT deficiencies be appropriately addressed.

Lawson et al[1] reported correlations between bias, precision, and total error as measured for laboratories participating in CAP Surveys and in the Great Lakes - Southeast Regional Quality Control Program, using the Quality Assurance Service (QAS) data program of CAP. Significant and moderate positive correlation of performance ranks was found for precision, bias, and total error for AST, glucose, and potassium, and of bias for phosphorus (Table 4). In addition, significant correlation was confirmed between quantitative survey and QAS bias for the four analytes, when analyzed by linear regression (Table 5). For AST, glucose, and potassium, QAS participants performed significantly better in surveys, with significantly lower bias, precision, and total error (Table 6).

Lum et al[2] in reporting on the relationship

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 QUANTITATIVE IMMUNOLOGY

YEAR	NON-LAP			LAP		
	NO.	UNACCEPTABLE(%)	ERROR	NO.	UNACCEPTABLE (%)	ERROR
1991	389	3.10	39.1	794	2.01	35.6**
1992	387	2.33	36.8	860	1.63	34.8*
1993	391	2.64	36.4	935	1.60	32.9**
1994	335	2.83	37.2	972	1.50	32.2**

Means significantly different by $p < .01$ (**) or $.01 < p < .10$ (*)

Table 3c

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 CATEGORICAL IMMUNOLOGY

YEAR	NON LAP		LAP	
	NO.	UNACCEPTABLE (%)	NO.	UNACCEPTABLE (%)
1991	2040	1.32	2008	0.96
1992	2057	1.59	2147	1.24
1993	2122	1.33	2337	0.96
1994	1786	1.01	2340	0.74

All mean differences significant $p < .01$

Table 3d

between laboratory performance in the Linearity Survey and that seen with concurrent PT, have documented a consistent and strong relationship between unacceptable survey results and calibration verification problems.[2] In addition, participants with performance-rated linear and verified calibration have lower rates of unacceptable results. Their study included some 33 analytes from the Chemistry, Ligand Assay, and Therapeutic Drug Surveys.

In studying consistency of PT performance during 1988-1990 and 1992-1994, two data sets lead to the same conclusion. The proportion of laboratories with consistent performance, i.e., 111,222,333 patterns, is greater than predicted with the blended hematology and immunology data from the earlier comparison study (Table 7) as well as within the latter specialty-specific study for chemistry, hematology, immunology, and

CORRELATION OF QAS & SURVEY DATA*					
ANALYTE	NO.	BIAS	ABSOLUTE BIAS	PRECISION	TOTAL ERROR
AST	88	.4893	.3256	.3573**	.4310
GLUCOSE	156	.7854	.4768	.3297	.4242
PHOSPHORUS	77	.5548	NS	NS	NS
POTASSIUM	64	.4206	.2570	.4758	.5505

*Spearman Correlation Coefficient
**N = 113
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 4

SUMMARY OF LINEAR REGRESSION RESULTS: CHEMISTRY SURVEY DATA-BIAS (y) vs GL/SE/NE QAS DATA (x)*				
ANALYTE	SLOPE	INTERCEPT,%	R	NO.
AST	0.70	0.14	.49	80
GLUCOSE	0.75	-0.83	.79	151
PHOSPHORUS	0.67	0.23	.60	72
POTASSIUM	0.64	1.20	.59	61

*From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 5

bacteriology (Table 8). In both the earlier and latter data sets, the ratio of observed to expected performance was higher for the 333 than for the 111 pattern. This suggests that within the set of consistent performers, relatively high performance is more difficult to sustain than relatively low performance. Observed consistency of performance, from two different 3 year study cycles, lends credibility to the PT process, by implying that performance is not random, but rather related to intrinsic operational characteristics of participating laboratories.

The effect of experience and time on chemistry, hematology, and immunology PT performance in the CAP EXCEL Survey has been reported by Tholen et al.[3] The data covered the 1987-1993 period. In the group as a whole, there is a tendency for progressive decrease in the average rates of unacceptable results with increasing years of participation (Table 9). Individual participant performance was also tracked. Significant improvement over time of participants was documented for all specialties (Table 10). Findings also suggest that laboratories with more experience with PT have higher rates of acceptable results

SURVEY PERFORMANCE vs QAS PARTICIPATION*				
ANALYTE	NOT IN QAS		IN QAS	
	NO. OF LABORATORIES	VALUE,%	NO. OF LABORATORIES	VALUE,%
AST				
Survey bias	5163	6.10	919	5.49†
Survey precision	5163	8.01	919	7.70‡
Survey total error	5163	10.41	919	9.69‡
GLUCOSE				
Survey bias	6632	3.18	1104	2.58‡
Survey precision	6632	4.10	1104	3.35‡
Survey total error	6632	5.36	1104	4.36‡
PHOSPHORUS				
Survey bias	3928	3.60	767	3.83
Survey precision	3928	3.94	767	4.13
Survey total error	3928	5.55	767	5.85
POTASSIUM				
Survey bias	6468	2.07	998	1.78†
Survey precision	6468	2.79	998	2.13‡
Survey total error	6468	3.55	998	2.86‡

*QAS indicates Quality Assurance Service; AST, aspartate aminotransferase
†P< .01 by Wilcoxon's test. ‡P < .0001 by Wilcoxon's test.
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 6

SURVEY QUARTILE PERFORMANCE OVER TIME - (1988-1990)
HEMATOLOGY & IMMUNOLOGY GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED N=2750

3 YEAR SEQUENCE	OBSERVED %	EXPECTED %	RATIO
111	4.3	1.6*	2.7
222	15.3	12.4*	1.2
333	6.4	1.6*	4.1

* $p < .001$ by x^2

1 = Highest Quartile
2 = Middle Two Quartiles
3 = Lowest Quartile

Table 7

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	CHEMISTRY, n=5017			HEMATOLOGY, n=1765		
	OBSERVED	EXPECTED	RATIO	OBSERVED	EXPECTED	RATIO
111	7.0	3.0*	2.3	16.8	10.9*	1.5
222	12.7	10.5*	1.2	2.7	2.4	1.1
333	5.1	0.8*	6.4	4.8	1.1*	4.4

* $p < .001$ by x^2
1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	IMMUNOLOGY, (Q) n=986			IMMUNOLOGY, © n=3480		
	OBSERVED	EXPECTED	RATIO	OBSERVED	EXPECTED	RATIO
111	23.1	19.6*	1.2	32.7	27.9*	1.2
222	1.5	0.7**	2.1	0.5	0.3***	1.7
333	2.8	1.0*	2.8	1.5	0.6*	2.5

* $p < .001$ by x^2
** $p < .01$ by x^2
*** $p < .05$ by x^2

1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8b

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	BACTERIOLOGY, n=4237		
	OBSERVED	EXPECTED	RATIO
111	12.7	6.7*	1.9
222	6.9	5.2*	1.3
333	4.7	0.9*	5.2

* $p < .001$ by x^2
1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8c

RATES OF UNACCEPTABLE RESULTS AND YEARS OF PARTICIPATION IN COLLEGE OF AMERICAN PATHOLOGISTS EXCEL SURVEYS, 1987-1993

SPECIALTY	No. of CHALLENGES	No. of Years of Participation 1	2	3	4	>4
Routine chemistry	1135	7.4	6.7	6.1	5.6	5.8
Therapeutic drug-monitoring Chemistry	200	6.8	7.3	7.8	6.6	5.9
Hematology						
Categorical	292	5.6	5.5	5.1	4.9	4.6
Quantitative	371	6.0	5.6	4.9	4.9	4.4
Common immunology	188	8.4	6.6	6.1	5.2	4.9
Special immunology	49	11.2	10.8	10.3	9.5	7.5
Blood bank	52	2.1	2.0	1.8	1.1	1.5

From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 9

PERFORMANCE IMPROVEMENT IN EXCEL SURVEYS*
TIME EFFECT p VALUES

	1987-1993		1989-1993		1991-1993	
	n	p	n	p	n	p
ROUTINE CHEMISTRY	247	<.001	632	<.001	1379	<.001
CATEGORICAL HEMATOLOGY	589	<.001	1236	<.001	2527	<.001
QUANTITATIVE HEMATOLOGY	612	<.001	1298	<.001	2668	<.001
COMMON IMMUNOLOGY	444	<.001	1009	<.001	2249	<.001

*Analysis of variance, experience vs. time effect
From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 10

PERFORMANCE IMPROVEMENT IN EXCEL SURVEYS
EXPERIENCE EFFECT p VALUES*

	1987-1993		1989-1993		1991-1993	
	n	p	n	p	n	p
ROUTINE CHEMISTRY	247	.585	632	.487	1379	.049
CATEGORICAL HEMATOLOGY	589	.813	1236	.275	2527	<.001
QUANTITATIVE HEMATOLOGY	612	.077	1298	.883	2668	<.001
COMMON IMMUNOLOGY	444	.065	1009	.002	2249	.015

*Analysis of variance, experience vs. time effect
From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 11

(Table 11). These results support the validity of PT by indicating that, as expected, prior experience and duration of participation are associated with performance improvements.

References

1. Lawson NS, Gilmore BF, Tholen DW. Multiprogram characterization of laboratory bias, precision, and total error. *Arch Pathol Lab Med.* 1988;112:454-461.

2. Lum G, Tholen DW, Floering DA. The usefulness of calibration verification and linearity surveys in predicting acceptable performance in graded proficiency tests. *Arch Pathol Lab Med.* 1995;119:401-408.

3. Tholen D, Lawson NS, Cohen T, Gilmore B. Proficiency test performance and experience with College of American Pathologists' Programs. *Arch Pathol Lab Med.* 1995;119:307-311.

4. Clinical Laboratory Improvement Amendments of 1988: final rule. *Federal Register.* Feb 28, 1992;55:7002-7186.

Performance in Proficiency Testing: An Indicator of Laboratory Quality?

Robert Rej, Ph.D.
Wadsworth Center for Laboratories and Research
New York State Department of Health
School of Public Health - State University of New York at Albany
Albany, New York

Abstract: Performance in external quality control or proficiency testing schemes is often cited as a measure of the quality of clinical laboratory testing throughout the world. There are significant differences which exist between routine testing of clinical specimens and samples tested for proficiency assessment. The demonstration of equivalent performance in proficiency testing and routine testing is a difficult association to establish. Differences in the mode of testing may include: requestor of laboratory services; characteristics of the specimens; sample transport; specimen identity to laboratory; pre-analytical variables; processing and accessioning; interferences and matrix effects; analytical phase; calculation of results; mechanism or reporting results; reference values; and application of test results. Only in the analytical phase and calculation of results are the processes nearly identical. It would be unexpected that performance in proficiency testing would be identical to routine performance unless these two phases contributed the largest source of error to the process. Nonetheless, split-specimen, or audit sample, testing for cholesterol and theophylline has demonstrated a significant correlation between routine performance and that based upon proficiency testing results.

Introduction

The ultimate objective of proficiency testing is the monitoring and improvement of health care through improving laboratory performance. Laboratory-improvement agencies typically rely on results of proficiency testing, along with on-site inspections, and regulations that specify educational requirements for staff for accreditation. There is no definitive information, however, describing which management attributes are of primary importance when related to performance on laboratory proficiency and which are of secondary importance.

In 1980, Peddecord and Cada[1] examined the effect of several variables on laboratory proficiency and concluded, at least for clinical chemistry and a few other branches of laboratory medicine, that enrollment in an external inspection and accreditation program is related to better performance. A 5-year review by the Centers for Disease Control and Prevention[2] confirmed that an overall program of inspection and accreditation generally improved laboratory performance over time. This review showed that the average number of major deficiencies (those which may have a direct effect on the quality of patient care or could affect the health and safety of hospital or laboratory personnel and must be corrected before accreditation can be extended to the laboratory) decreased from 16 to 6 over the 5-year period.

Improvement in laboratories cannot always be measured in objective or direct terms since factors intermingle and overlap to the point that it would be inappropriate to suggest that laboratory improvement is

solely due to the analytical proficiency testing component of the accreditation process. Several empirical studies, however, have suggested that continued participation in proficiency testing programs is related to improved performance.[3-10]

The Clinical Laboratory Improvement Amendments of 1988 (CLIA'88) have helped the laboratory community in the United States to renew interest in defining the true role of proficiency testing. The CLIA'88 legislation itself calls for assessing of "validity, reliability, and accuracy of proficiency testing."[11] This is a charge to evaluate the effectiveness of proficiency testing. Several questions, while straight-forward at first, are rather complex and difficult to address. Does accuracy of proficiency testing entail that results are exact predictors of those that would be obtained on patient testing or that results are correlated with the quality of patient testing? This review examines the strengths and limitations of proficiency testing as an evaluator of laboratory performance.

Differences in the Process: Patient vs. Proficiency Testing

The clinical laboratory testing process is comprised of several phases and components. One classification scheme may be: the requestor of laboratory services; characteristics of the specimens; sample transport; specimen identity to laboratory; pre-analytical variables; processing and accessioning; interferences and matrix effects; analytical phase; calculation of results; mechanism or reporting results; reference values; and application of test results. This is shown schematically in Table 1, with some potential differences and similarities between the proficiency- and the patient-testing processes. The highlighted area note the two phases where, in my view, considerable similarity and overlap exist. Key areas are reviewed here:

Analytical specificity, interferences, and matrix effects

Are the specimens used in proficiency testing similar to authentic patient samples encountered in routine laboratory testing and therefore a realistic challenge of performance? It is known that so-called matrix effects may give rise to artificially induced errors in proficiency testing. Methods that are exquisitely sensitive to matrix effects, however, are similarly sensitive to alterations in patient sera, and such factors can be assessed.[12-17]

Lyophilization may introduce errors not normally encountered with processing patient specimens.[12] To examine the extent of proficiency test specimen matrix effects, we distributed three specimens differing primarily in their matrix composition in a single proficiency testing event of laboratories enrolled in the New York State survey for clinical chemistry.[15] These were: pooled normally clotted liquid human serum with minimal supplementation, liquid serum prepared by re-calcification of pooled human plasma, and commercially lyophilized serum prepared by re-calcification of pooled human plasma. Wherever possible, analyte concentrations were adjusted to be comparable in all three specimens. Twelve chemistry analytes were selected for comparison including lipid, enzyme, substrate, and ionic constituents. Significant differences in the inter-method behavior (commutability) were found amongst the three types of specimens for HDL-cholesterol with inter-laboratory coefficients of variation (CV) of 11.4% (liquid serum), 28.7% (liquid re-calcified

Specimen Type:	**A Proficiency Sample**	**A Patient Sample**
Requestor:	Ordered by HCFA, proficiency test provider	Ordered by physician or health provider
Sample Characteristics:	Sample obtained from large pool	Client sample obtained from individual
Sample Transport	Transport in mail;	Transport within/among institution(s);
Specimen Identity to Laboratory:	Usually identified; Unique vial or tube	Relatively anonymous - usually one of many
Preanalytical Variables:	Reconstitution errors	Patient preparation; specimen collection; sample collection device; sample pretreatment and centrifugation
Entry into Process:	Enter process at a later stage	Enter process at earliest stage
Accession:	May require special accessioning to avoid creating patient record	Usually routine
Interferences:	Matrix effects due to lyophilization or preparation not seen with patients	Drugs and metabolite effects usually not seen with proficiency specimens
Analysis:	Should be routine; may require special handling due to sample characteristics or analyte level	Usually routine; may require special handling due to analyte level
Calculation of Results:	Should be routine; may require special calculation due to dilution of specimen	Usually may require special calculation due to dilution of specimen
Mechanism or Reporting Results:	Extraordinary reporting (usually manual)	Routine reporting (usually electronic)
Reference Values:	May differ amongst laboratories	Usually uniform
Application of test results:	Result used for laboratory evaluation and/or accreditation	Result used for patient care

Table 1. Differences and Similarities between the Proficiency Testing Process and Routine Clinical Laboratory Analyses.

plasma), and 52.1% (lyophilized). For the other 11 analytes, liquid serum and liquid re-calcified plasma demonstrated similar commutability, while in nearly each case lyophilization introduced considerable matrix effects. With a specimen of liquid origin, a single normal distribution was found for total creatine kinase (mean = 149 U/L, CV = 11.2%, Figure 1 shaded bars), while an apparent bimodal distribution was observed for a lyophilized material using identical analytical methods (mean= 122 U/L, CV = 33.1% Figure 1 open bars). Inter-laboratory coefficients of variation were considerably larger with the lyophilized material for most, but not all analytes, indicating that errors in reconstitution/filling were not the predominant source of variation. CVs (%) were liquid and lyophilized materials were, respectively: glucose 6.9, 6.9; sodium 2.0, 2.0; chloride 3.0, 5.5; cholesterol 3.9, 5.9; creatinine 10.4, 44.9; calcium 3.6, 6.9.

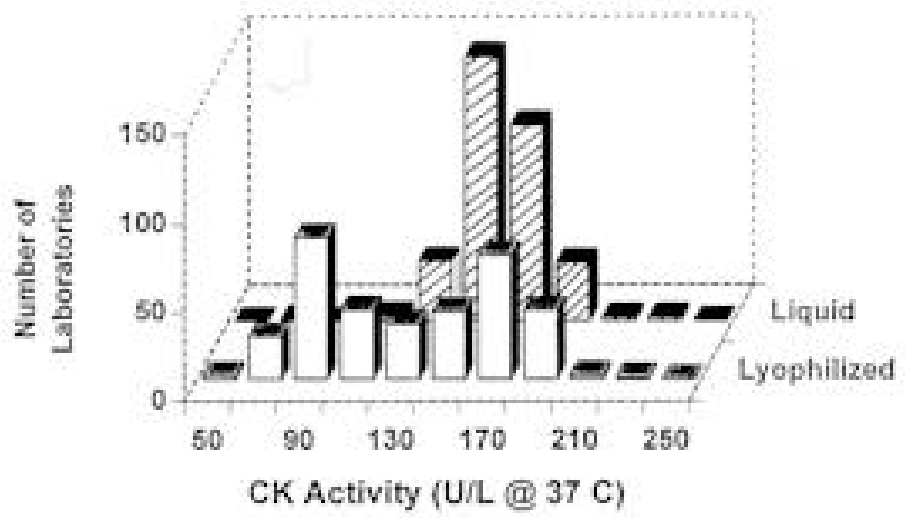

Figure 1. CK Activity by Number of Laboratories, liquid and lyophilized specimens.

There is also the potential for a "reverse matrix effect," whereby interferences in some authentic patient specimens (metabolites and/or drugs) are absent from proficiency specimens, and proficiency testing samples do not provide the range of interferences encountered in routine analysis.[18]

The Analytical Phase

This is the focal point of the clinical laboratory process and one where the proficiency and patient specimen can undergo identical processing. Although this would be ideal, a survey of laboratorians in hematology and clinical chemistry suggests that some special treatment of proficiency test specimens was commonplace a decade ago[19]. It has also been observed that special treatment of external quality control specimens can result in improved performance.[20]

In reviewing proficiency test records and laboratory inspections over the past two decades, it has been my experience that “special treatment” can also lead to poorer performance in a proficiency survey due to the fact that it is out-of-the-routine. The entire range of analyte concentration will inevitably be larger in the proficiency testing specimens for most analytes since these can usually be supplemented at concentrations far outside physiological ranges.

	First Testing	Second Testing
Average Score (LH)	80.1%	91.4%
Number of Laboratories Failing (LH)	19	6
Average Score (FSH)	85.3%	96.6%
Number of Laboratories Failing (FSH)	10	1

Table 2. Performance of Participant Laboratories Upon Test Introduction for LH and FSH in the New York State Endocrinology Proficiency Testing Program

Furthermore, differences amongst the type of laboratories will also affect distribution of analyte concentration (with smaller variations being observed in large reference and university hospitals while larger variations being observed in physician office laboratories). Substandard performance in a proficiency test at extremes of analyte ranges will not provide data that allow projection to the performance likely to be found within usual reference values. Special handling (dilution of samples with elevated concentrations of analyte) or method of presentation to the instrument (e.g., syringe injection or aspiration of blood gas specimens) may be required for proficiency test specimens.

Reference values and application of test results

Although uniform criteria of evaluation are provided by approved CLIA'88 proficiencytesting providers, use of peer group evaluation may perpetuate use of procedures that cause personnel to perform in a clinically unacceptable manner. Although evaluation by peer groups should be performed only for specimens that demonstrate a "matrix effect," criteria for establishing peer groups are vague, and subtle interactions between a method group and a given proficiency testing material may well be treated differently amongst proficiency testing providers. Data on the mechanism used to establish a target value should be available, and if overall participant mean, peer-group mean, or reference method value was used in establishing the target.

Influence of the Proficiency Testing Process Itself

Proficiency testing is usually not a passive barometer that merely monitors laboratory performance. Participating in a proficiency testing program is interactive and genuine poor performance is examined and corrected in most laboratories. The actual performance of laboratories not involved in an external quality assessment scheme is difficult to estimate. This might be gauged, however, by examining the performance of laboratories that are newly enrolled in a

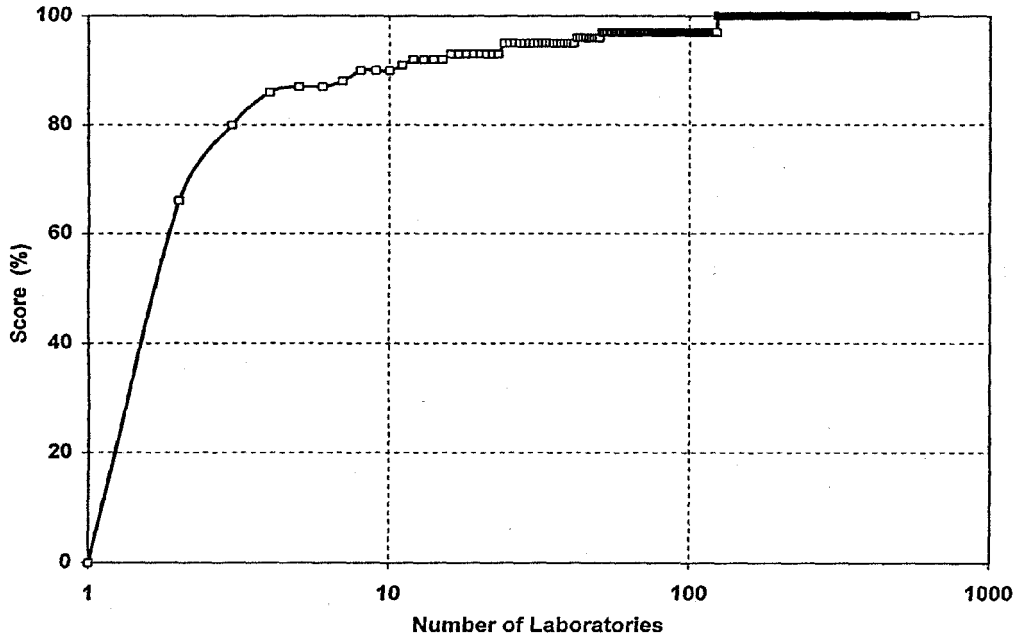

Figure 2. Improvement in proficiency scores by laboratories.

proficiency testing program or with established laboratories when a new analyte is introduced. Table 2 demonstrates that in the 4-month period intervening between the first and second testing in the New York State proficiency testing program, a dramatic improvement in performance can be found immediately after the introduction of proficiency testing for lutropin (LH) and follitropin (FSH). This improvement in performance, both in improved average scores in proficiency tests and reduction in numbers of failing laboratories, was due to two factors: voluntary withdrawal of testing for these analytes by some laboratories and improved performance by those remaining in the program.

Improvement in performance as evidenced by analysis of laboratory proficiency testing results has been demonstrated in the Regional Quality Assurance Programs in the U.S., elsewhere in North America, and throughout the world. Accordingly, examination of laboratories regularly participating in proficiency testing for several proficiency cycles will likely result in examination of reasonably fine distinctions amongst laboratories.[21] This is demonstrated in
Figure 2. Of approximately 600 laboratories participating in the New York State Hematology proficiency test, 10 laboratories failed to achieve scores > 90% (using the CLIA'88 grading schemes), and most (78%) attained a score of 100%. Only two laboratories failed to achieve an overall passing score of 80%.

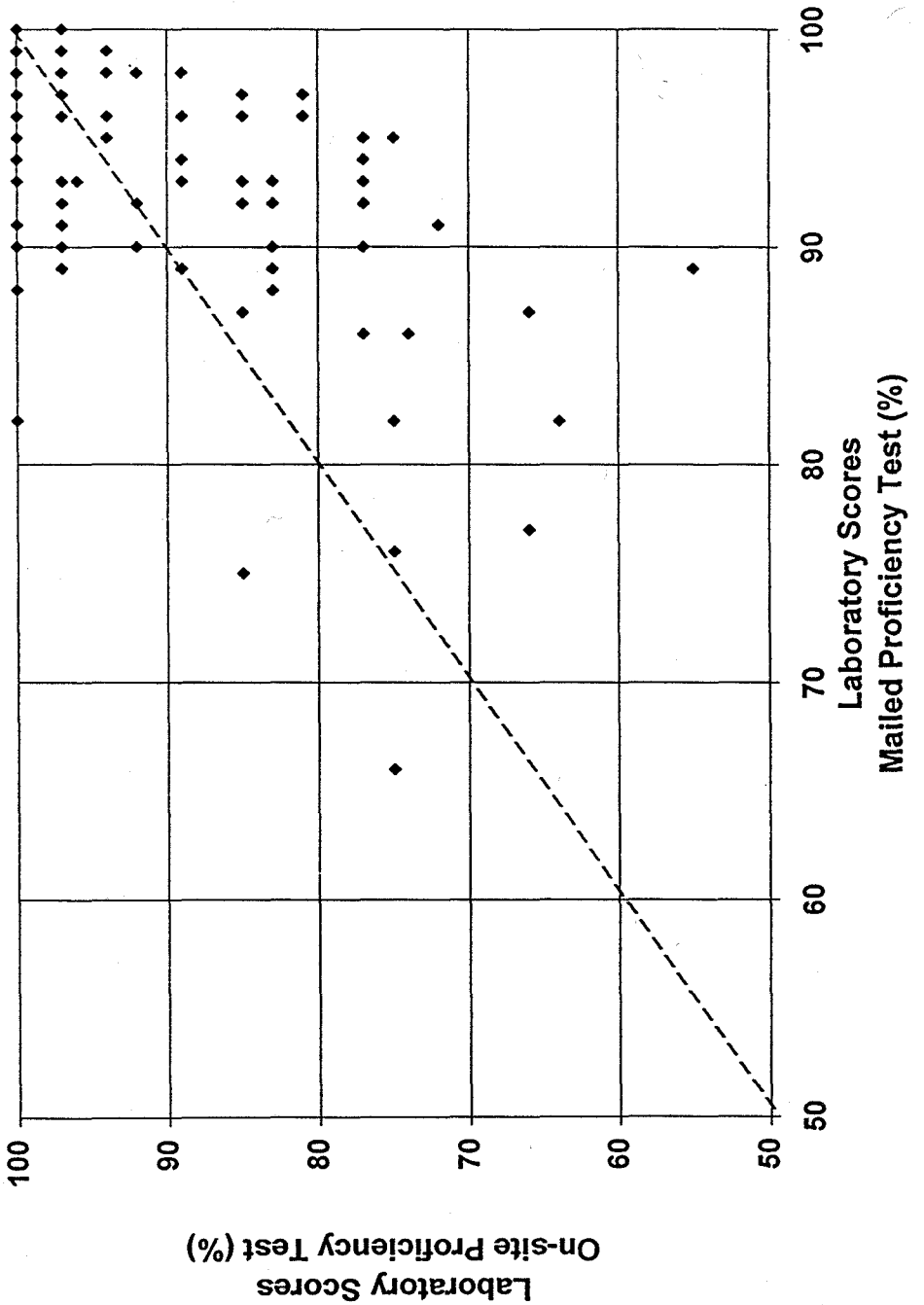

Figure 3. Comparison of laboratory scores, mailed proficiency testing vs. on-site proficiency testing.

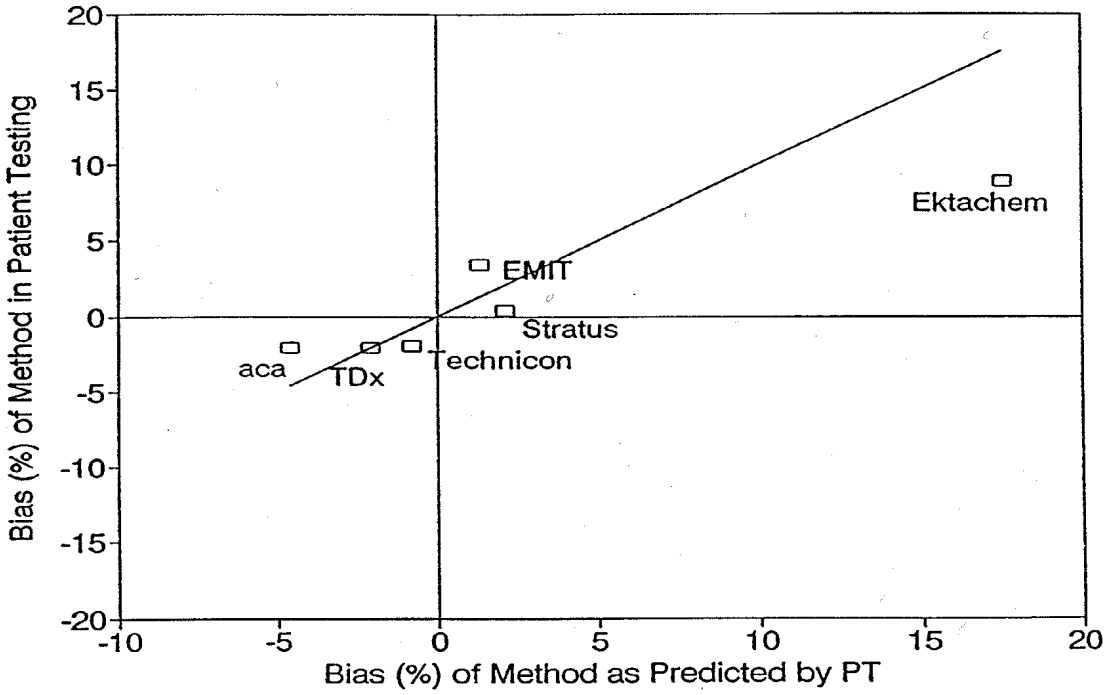

Figure 4. Comparison of bias (%) of method predicted by PT vs. bias (%) of method in patient testing.

Auditing Proficiency Testing

A number of mechanisms have been devised as audits of results oabtained by conventional proficiency testing. Three are reviewed here:

"Blind Submission" of Proficiency Samples

In this schemem, samples used in proficiency tests are distributed to laboratories disguised as "patient" specimens.[22] Although this mechanism may circumvent some special treatment of proficiency test samples, it may not provide the same information, because even if they are treated in an identical fashion, pre-analytical and post-analytical processing may differ (Table 1).

Overt proficiency testing samples enter the system at the analytical phase and are subject to extraordinary reporting, whereas blind proficiency testing samples enter the system at an early phase and are subject to routine reporting.

On-site Proficiency Testing

Some information may be gained from examining routine proficiency testing distributed by scheduled mailing and that presented to laboratories during inspection. In the area of blood pH and gases, the proficiency test program organized by the Wadsworth Center presented specimens in both manners; four sets (3 vials) by scheduled mail; one set (3 vials) presented at the time of unannounced inspection. Blood

gas measurements may represent an analysis where special treatment (increased calibration, replacement of electrodes, etc.) can be effected for routine proficiency testing but not possible at the time of inspection. Results are shown in Figure 3. A high degree of correlation was found between the results of the scores obtained (r^2 = 0.49). A slight, but statistically significant (P< 0.01) by paired t-test), difference was observed between scores obtained by on-site (mean = 87.6%) and mailed (mean = 91.6%) testing routes.

Split-specimen patient testing

To better examine routine laboratory performance, at the time of annual laboratory inspections conducted by the New York State Health Department, we obtained aliquots of each of two sera that had been analyzed for cholesterol or theophylline by the inspected laboratory.[23,24] These aliquots were mailed to our laboratory; we also obtained the clinical results determined and reported for those specimens by the laboratory. Specimens were stored at < -60° C and analyzed by reference methods (CDC modified Abell-Kendall for cholesterol and HPLC for theophylline). Results were obtained for > 200 laboratories. We found that the predictive value of proficiency testing performance in assessing quality of routine testing was high; for theophylline, 100% for predicting substandard reliability of routine patient testing and 94% for excluding substandard reliability of patient testing. Significant correlation was found between analytical bias observed in proficiency tests and that found for patient testing (Fig. 4). For cholesterol, the average difference between participant performance and the reference method was a positive bias of 1.47%; this is equivalent to the overall population bias measured by our routine proficiency testing. Most reported results (88%) were within $\pm$ 10% of the value determined by the reference method. This is similar to performance determined by our proficiency program, where ca. 15 % of results were beyond $\pm$10% of the reference method target value. Using NIH guidelines for risk assessment (200 and 240 mg/dL), 13 specimens (4.4%) were misclassified to a lower risk; 7 specimens (2.4%) were misclassified to a category of higher risk.

We found this manner of auditing laboratory performance effective in that true patient specimen results and the results reported are used in the evaluation process. A similar split- specimen testing study is under way for calcium analysis, using atomic absorption as the reference technique. This analyte meets many of the criteria shared by cholesterol and theophylline (stability, availability of reference methods, a wide variety of analytical procedures) and is an analyte where analytical goals are stricter that current performance ability.

References

1. Peddecord KM, Cada RL. Clinical laboratory proficiency test performance. Its relationship to structural, process and environmental variables. *Am J Clin Pathol*. 1980;73:380-5.

2. Duckworth JK. Voluntary inspection and accreditation and improvement of laboratory performance. The private sector view. In: Quality Assurance in Health Care: A Critical Appraisal of Clinical Chemistry. RN Rand, RJ Eilers, NS

Lawson, A Broughton (Eds.), American Association for Clinical Chemistry, Washington, DC, 1980, pp. 381-90.

3. Copeland BE, Rosenbaum JM. Organization, planning and results of the Massachusetts Society of Pathologists regional quality control program. *Am J Clin Pathol.* 1972;57:676-88.

4. Vincent WF. The Perspective of Connecticut Department of Health. In: Proceeding of Second National Conference on Proficiency Testing. M Davis (Ed.), Information Services, Inc., Bethesda, MD, 1975, p 18.

5. Pinkerton PH, Wood DE, Burnie KL, et al. Proficiency testing in immunohematology in Ontario, Canada, 1975-1977. *Am J Clin Pathol.* 1979;72:559-63.

6. Shephard MDS, Penberthy LA, Fraser CG. Evolution of a national urine quality assurance programme: the Australasian experience, 1981-1983. *J Clin Pathol.* 1984;37:415-23.

7. De Leenheer AP, Thienpont LMR. External quality assessment of laboratory performance in haematology in Belgium: analysis of two and a half year's experience. *Clin Chim Acta*. 1984;144:95-103.

8. Whitehead TP, Woodford FP. External quality assessment of clinical laboratories in the United Kingdom. *J Clin Pathol.* 1981;34:947-57.

9. Jansen RTP, Jansen ADP. A coupled external/internal quality control program for clinical laboratories in The Netherlands. *Clin Chim Acta.* 1980;107:185-201.

10. Hill PG, Kanagasabapathy AS. Improvement in laboratory performance by an interlaboratory quality control programme. *Indian J Med Res.* 1979;6:853-64.

11. Public Law 100-578 (HR 5471) Clinical Laboratory Improvement Amendments of 1988; Section 4-a-1. Congressional Record 134, October 31, 1988.

12. Rej R, Jenny RW, Bretaudiere JP. Quality control in clinical chemistry: characterization of reference materials. *Talanta* 1984;31:851-62.

13. Rej R. Accurate enzyme activity measurements. Two decades of development in the commutability of enzyme quality control materials. Arch *Pathol Lab Med.* 1993;117:352-64.

14. Rej R. Proficiency testing, matrix effects, and method evaluation. *Clin Chem*. 1994; 40:345-6 [Editorial]

15. Rej R, Norton-Wenzel C. Matrix effects and external assessment of laboratory measurements using liquid and lyophilized specimens. *Clin Chem*. 1994;40:1000 [Abstract].

16. Bretaudiere JP, Rej R, Drake P, Vassault A, Bailly M. Suitability of control materials for determination of alpha-amylase activity. *Clin Chem*. 1981;27:806-15.

17. Bretaudiere JP, Dumont G, Rej R, Bailly M. Suitability of control materials. General principles and methods of investigation. *Clin Chem*. 1981;27:798-805.

18. Jenny RW. Interlaboratory evaluation of salicylate interference in colorimetric acetaminophen methods and its clinical significance. *Clin Chem*. 1985;31:1158-62.

19. Rowan RM, Laker MF, Alberti KG. The implications of assaying external quality control sera under "special conditions." *Ann Clin Biochem*. 1984;21:64-8.

20. Cembrowski GS, Vanderlinde RE. Survey of special practices associated with College of American Pathologists proficiency testing in the Commonwealth of Pennsylvania. *Arch Pathol Lab Med*. 1988;112:374-6.

21. Rej R, Jenny RW. How good are clinical laboratories? An assessment of current performance. *Clin Chem*. 1992;38:1210-17; discussion 1218-22.

22. Hansen HJ, Caudill SP, Boone DJ. Crisis in drug testing. Results of CDC blind study. *JAMA* 1985;253:2382-87.

23. Jenny RW, Jackson KY. Proficiency test performance as a predictor of accuracy of routine patient testing for theophylline. *Clin Chem*. 1993;39:76-81.

24. Rej R, Norton CS. External assessment of laboratory cholesterol measurements using patient specimens. *Clin Chem*. 1989;35:1069 [Abstract].

Achieving and Assessing Acceptable Analytical Performance: The Challenge of Matrix Effects

Fred D. Lasky, Ph.D.*
Stephen G. Daggett, Ph.D.
Johnson & Johnson Clinical Diagnostics, Inc.
Rochester, New York

***Presenting Author**

Abstract: Accurate and reliable routine *in vitro* diagnostic testing is needed by physicians to help provide appropriate care for their patients. These goals, which have been common to both health care professionals and manufacturers, are now mandated by law. Achieving and assessing acceptable performance have a number of common challenges. Once acceptable levels of accuracy are defined, performance can be evaluated by component: precision; bias; specificity (random interferences); and stability. Each component affects our ability to achieve and maintain performance within an acceptable window. A key element in achieving and assessing the accuracy of test results is the use of well characterized reference or designated comparison methods that are performed under stringent process control. A feature often overlooked in the analytical process is the specification of sample matrices to be analyzed. Differences in composition between patient samples and processed fluids present a challenge for calibrating and assessing performance of routine methods. To manage these differences in sample matrices, we have demonstrated success in establishing calibration of routine methods with patient samples, so that routine methods correlate with designated methods. From these correlations, values are assigned to calibrators. Effective process control contributes to vial-to-vial uniformity. Calibration fluid stability is enhanced by saccharide stabilizers that displace outer-sphere protein-bound water, which aids effective lyophilization. Calibrator set points can be established efficiently when analyte concentrations or activities are adequately recovered after reconstitution of lyophilate. Compatibility of components (analytes, stabilizers and additives) is also desired to prepare economic, multi-purpose fluids. This is a significant challenge that also faces proficiency testing providers, whose fluids are similar in preparation and composition. Alternative strategies to traditional proficiency testing schemes (using lyophilized fluids) are achieving good success. Individual or pooled patient samples may better demonstrate method performance at clinically significant concentrations, although they do increases risk of biohazard exposure. This approach is also consistent with manufacturers' efforts to develop methods that perform well with patient samples. Although processed fluids may provide an adequate tool to assess consistency of results across laboratories for similar methods and instruments, assessing accuracy will continue to require the use of patient samples.

Introduction

Accurate and reliable *in vitro* diagnostic testing is needed by physicians to help provide appropriate care for their patients. These goals are now mandated by law! Achieving and assessing acceptable levels of accuracy have several common challenges, especially when artificial, processed fluids

			Source of Variability	
Analyte	Target	Total C.V.	Lyophilizer	Vial & Rep
CHOL	139.3 mg/dL	1.58%	1.93%	98.07%
GLU	90.9 mg/dL	0.69%	6.12%	93.88%
Na^+	122.9 mmol/L	0.44%	2.79%	97.20%
ALKP	98.8 U/L	0.94%	13.81%	86.18%
CK	168.7 U/L	2.33%	7.43%	92.56%

Table 1:Lyophilizer Qualification: Variance Component Analysis

are used. Although there is no single, agreed-upon standard, acceptable levels of accuracy must be established for meaningful evaluation. To better understand and control the total allowable error (or acceptable accuracy), we use a model in which error components are identified as precision, bias, specificity (random interferences) and stability.[1]

Desirable characteristics of processed fluids include uniformity, stability, analyte recovery and compatibility with the reagent and instrument. When used as calibrators, processed fluids affect bias, laboratory-to-laboratory precision and stability. The same characteristics affect the perception of system performance when they are used in proficiency testing (PT) programs.

Achieving Accuracy

A key element in achieving the accuracy of test results is the use of well characterized reference or designated comparison methods that are performed under stringent process control, such as described in the ISO 25 Guidance for *General Requirements for the Competence of Calibration and Testing Laboratories.* A feature often overlooked in the analytical process is specification of sample matrices to be analyzed. Additionally, performance limits for reference methods must be established and maintained at significantly more stringent levels than what is expected of routine methods.[2,3] International Reference Preparations (IRP) may be used where reference methods are not available or are unlikely to be developed. Differences in matrices (matrix effects) between IRP and patient samples as well as between IRP batches present an additional challenge in calibration.

We have demonstrated success in maintaining acceptable levels of accuracy in routine methods by establishing calibration through correlation with our designated methods using patient samples.[4] Using these correlations, values are assigned to calibrators to transfer comparable performance from the factory to the field. Calibrator properties, therefore, affect performance.

Fluid Manufacturing; Process Challenges

Manufacturing fluids that meet the requirements of calibrators is a challenge that requires careful product design and process control. Some characteristics, such as uniformity, clarity and reconstitution time, are related to controlling the lyophilization process. We demonstrated that our process is capable of acceptable vial-to-vial uniformity, with volumetric transfers and analyte measurements contributing more

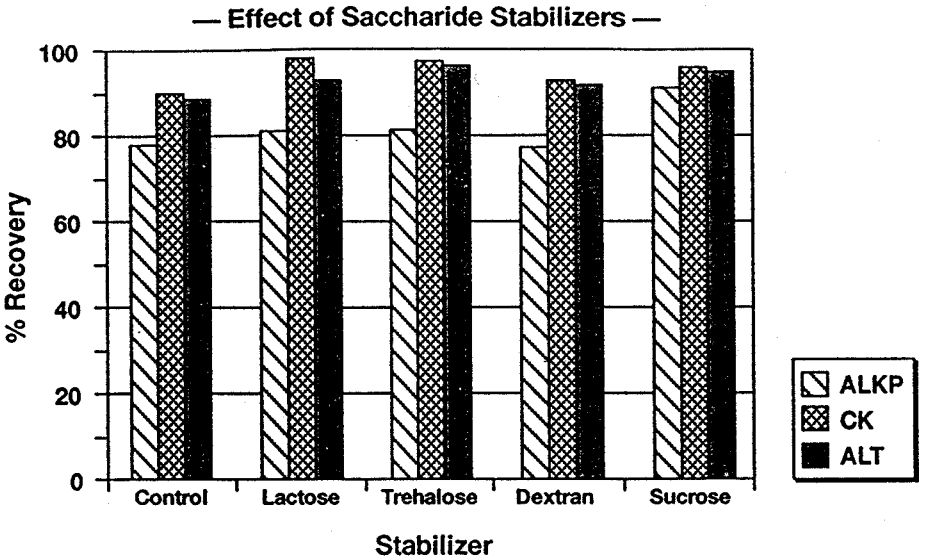

Figure 1. Enzyme Recovery in Reconstituted Fluid

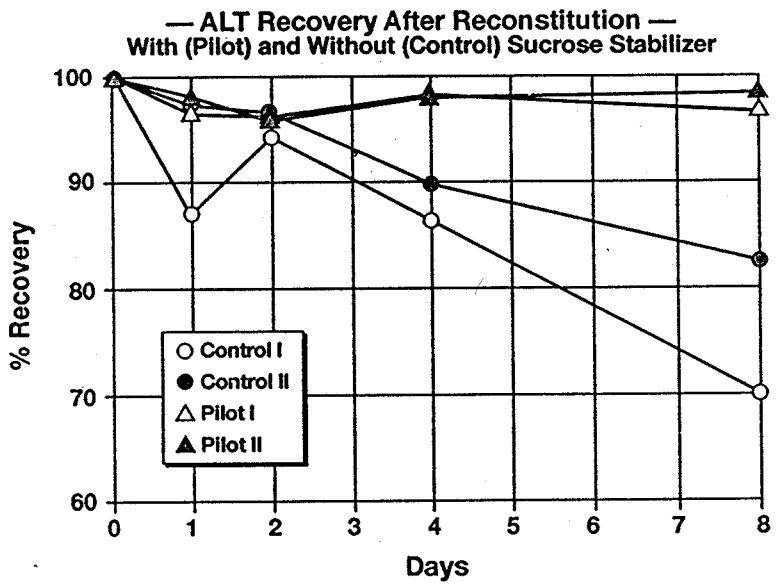

Figure 2. Lyophilate Stability at 50°C

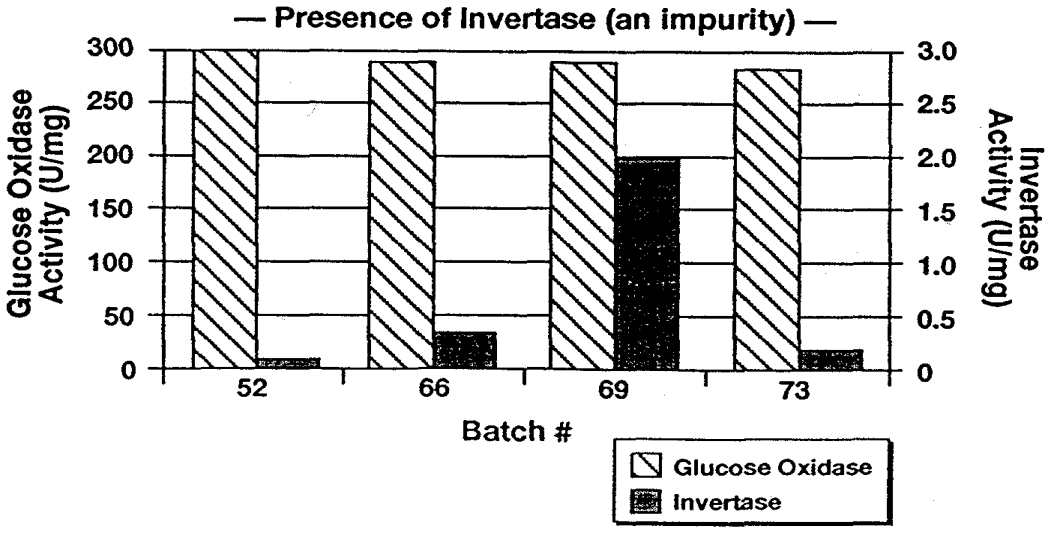

Figure 3. Glucose Oxidase Batch Analysis

variability than lyophilization (Table 1).[5] As expected, protein analytes are more sensitive to manipulation and processing.

Caution must be used when supplements are added, e.g., analytes, stabilizers and additives, to prepare economical, multi-purpose fluids. During development of an enzyme calibrator, unexpected inhibition of CK was observed when amylase was added to the pilot mix. CK activity dropped from 620 U/L (control) to 35 U/L when amylase was added. An alternative supplier's material was satisfactory; CK = 622 U/L. Porcine was the source for all enzymes, heart for CK and pancreas for amylase.

Calibrator stability is enhanced by adding saccharides, which displace outer-sphere protein-bound water and reduce collapse of the cake during lyophilization. Sucrose was selected as a candidate because of better enzyme recovery after reconstitution (Figure 1). Furthermore, lyophilate-enhanced stability was observed in an accelerated storage test at 50°C (Figure 2).

Significant variability, however, was noted between reagent lots of glucose slides using a pilot calibrator preparation (differences up to 50 mg/dL at 130 mg glucose/dL). Raw material batch analysis of glucose oxidase (GO), the active reagent, determined the presence of invertase, an impurity in GO, varying by batch (Figure 3). Invertase converts sucrose to glucose and fructose, thus causing an artifactual increase in the amount of substrate measured.

These same challenges face PT providers, whose fluids are similar in manufacture, preparation and composition to our calibrators. These experiences demonstrate some potential pitfalls in validating the

			Lyophilized			Fresh Serum			
Analyte	Units	AMM*	Interval	AMM	Interval	AMM	Interval	AMM	Interval
Cl^-	mmol/L	99	15	125	16	101	9	102	10
Na^+	mmol/L	134	15	155	16	141	10	141	10
Creatinine	µmol/L	89	75	280	75	89	30	188	50

*AMM=All Method Mean

Table 2. Interval Covering 95% of Participants' Results

	Cholesterol				HDL-Cholesterol	
Sample:	1	2	3	4	1	2
NY State Wadsworth Center	5.67	6.47	4.82	6.14	1.44	1.17
AMM	5.61	6.47	4.75	6.27	1.41	1.15

Table 3: Participants' Results for Cholesterol and HDL-cholesterol (mmol/L)

compatibility of fluid components and formulations with reagent lot composition.

Assessing Accuracy

Several strategies have been attempted to better achieve PT's objective, which is to provide a measure of test system reliability and accuracy. CAP evaluated lyophilized bovine and human serum matrices. Protein-based analytes showed no decrease in variability in the human matrix, while human serum occasionally had greater variability than bovine.[6] We had similar experiences where the matrix (human, bovine or goat serum, or BSA) has less effect on obtaining desirable fluid characteristics than effective control of manufacturing processes.

Fresh human samples (with and without supplementation) have been tried by several PT providers with good success. CAP has a study ongoing; the Veterans Administration (VA) has a program in which both lyophilate and fresh samples are used. Ontario's Laboratory Proficiency Testing Program has published results showing that consistency was significantly better for fresh serum than for lyophilate. [7](see Table 2)

The all-method means for fresh serum cholesterol and HDL cholesterol were remarkably close to results from a CDC-network reference laboratory. (see Table 3)

Additionally, all methods demonstrated cholesterol performance with biases less than 1.5% which was well within the NCEP goal of 3%.

Patient samples better demonstrate method performance at clinically significant concentrations, although they do increase the risk of biohazard exposure.

Improving Program Utility

Using patient samples in PT programs is also consistent with manufacturers' efforts to develop methods that perform well with clinical samples. System changes designed to improve performance with patient samples might, by serendipity, also result in better PT performance. For example, method-specific

means moved closer to the all-method mean when a manufacturer enhanced the on-analyzer stability of its phosphorus method; another changed the read wavelength to improve analyzer model-to-model consistency for glucose. In both cases, however, the objective was to improve method operational characteristics or performance.

From time to time, changes are made in PT fluid manufacture to better assess performance of a method or method group. We worked with CAP to include bicarbonate diluents. A 10% negative bias was eliminated from our urea results when the broad physiologically expected range of CO_2 was present. Hitachi's diluted Cl^- ISE method also improved.

Improving, validating and controlling changes in test systems for and with processed fluids is a daunting task. Reagent changes (suppliers, raw materials, process improvements), combined with changes in fluid batches, matrices (from program to program) and the variety of component additions, make the number of independent variables that must be controlled unmanageable. Differences in results between processed and "native" serum are not surprising when test systems are optimized for use with patient samples; a frequent reminder that defining the appropriate sample -- patient samples -- is a fundamental principle of metrology.

Quality Opportunities

To determine where opportunities lie, we surveyed 17 clinical chemists with the following question: What are the five most important quality issues that are associated with (1) overall clinical laboratory operations (from test request to result utilization), and (2) are broadly related to the analytical services provided by your laboratory which would most benefit the patient if resolved now?

Analytical element concerns were only 13% of the responses, while pre- and post-analytical opportunities were noted 40% and 46%, respectively. Within the laboratory, responses were more diffuse. Of 78 responses, the top issues (with frequency) were: Equivalent results across methods (12); Enhance reagent stability & reduce variability (10); and Improve personnel training & competency. Improved PT was far down the list, mentioned only twice.

We asked representatives of five major manufacturers of chemistry systems about the objectives of their improvement and development programs. Their responses were consistent: Such programs are aimed to improve performance with patient specimens. Not one manufacturer could recall an improvement program being initiated solely because of PT results.

Observations

Because processed fluid manufacturing is so dependent on external factors, such as matrix and variable attempts to mimic the physiologic composition of human serum, we believe its use in assessing method accuracy is fraught with insurmountable limitations. Despite this, PT continues to provide a good assessment of laboratory-to-laboratory *consistency* for total test systems.

Fresh specimens provide better assessments of repeatability and accuracy because methods are designed for these samples. Additionally, assessing accuracy at the clinically important concentrations is manageable with the use of fresh patient samples.

Finally, the benefits of any improvements in reagent specificity or fluid processing to

enable traditional PT programs, which use lyophilized materials, to better assess method accuracy and reliability must be weighed against the costs of diverting resources needed to develop new tests. We surveyed five leading manufacturers of clinical systems, and *none* has ever directed improvement efforts to anything but patient performance. Not surprising, especially now that health care priorities are being critically scrutinized!

Recommendations

Programs are needed to assess the reliability of laboratory tests. With limited laboratory resources, however, priorities must be established *analyte by analyte* to determine which we deal with first. Only then should limits be established for acceptable repeatability and accuracy. Then, costs of establishing a Reference Laboratory Network for the critical analytes need to be determined. This network should be international in scope; include manufacturers, which have resources and often special expertise; require stringent process and procedural control; and require participating laboratory performance to be *significantly better* than routine methods. Determine the effectiveness of alternative strategies, such as having the manufacturer verify *accuracy* of its systems through the network asfrequently as necessary, while clinical laboratories continue to verify *consistency* across laboratories for similar methods and instruments through traditional PT programs. Occasionally, fresh samples (or pools) should be included in PT programs to verify accuracy, especially at critical concentrations. Fresh samples can be relied upon by the laboratory, manufacturer, and government agencies to assess results on the same samples for which the systems are designed and on which the physician depends for patient evaluation.

References

1. Lawton WH, Sylvestre EA, Young-Ferraro BJ. Statistical comparison of multiple analytic procedures: Application to clinical chemistry. *Technometrics* 1979:21;397-409.

2. Bennett ST, Eckfeldt JH, Belcher JD, Connelly DP. Certification of cholesterol measurements by the National Reference Method Network with routine clinical specimens: Effect of network laboratory bias and imprecision. *Clin Chem*. 1992:38;651-657.

3. Fricker RF, Ritter M, Lasky FD, Greenberg N. Comparison of alternative statistical approaches to external quality assessment of cholesterol performance. *Clin Chem*. 1992;38:1028 [Abstract]

4. Lasky FD. Achieving accuracy for routine clinical chemistry methods by using patient specimen correlations to assign calibrator values: A means of managing matrix effects. *Arch Pathol Lab Med*. 1993;117:412-419.

5. Henderson RC, Peters C. Performance of calibrator and control fluids manufactured using high performance lyophilizers. *Clin Chem*. 1995:41;S215 [Abstract]

6. Arkin C. Man v. beast: Comparing human and bovine QC materials. *CAP Today*. 1994;8:49-51.

7. Ontario Laboratory Proficiency Testing Program (LPTP). Review of activities.

1994.

8. Laboratory Proficiency Testing Program. Review of activities. Toronto; LPTP, 1994:36.

Proficiency Testing 1995-2000: Educational Tool, Quality Control Tool, Management Tool, or a Regulatory Tool?

Sharon S. Ehrmeyer, Ph.D.*
Ronald H. Laessig, Ph.D.
Department of Pathology and Laboratory Medicine
University of Wisconsin-Madison
Madison, Wisconsin

***Presenting Author**

Abstract: Proficiency testing (PT) is all of the above; numerous citations demonstrate the efficacy of each application. The real question is: "What is PT to be, circa 2000, in the context of CLIA'88?" PT was conceived by and for laboratory directors as an educational tool based on interlaboratory performance comparisons. As a quality control (QC) tool, PT has not reached its potential. The drawback is the lack of timeliness in reporting. More recently, with mandated director's review of responses to PT failures, it is an integral part of management and quality assurance. In 1995, PT is the lynchpin of the CLIA'88 regulatory process. The data, from laboratories participating in PT for the first time in 1994, demonstrate both its ability to affect performance improvement, as well as, PT's known limitations. The opportunity in the future for PT depends on redefining its mission and role. To transform PT to a new level of effectiveness as a QC tool, timeliness must be addressed. Minimizing turnaround-times (TAT) through technology will allow a quantum improvement in the value of the PT process. If the 60 day TAT, currently mandated under CLIA'88, becomes 60 seconds or less, a whole new quality paradigm is possible. The challenge of implementing this new vision is only to dare to dream!

Introduction

Interlaboratory proficiency testing has nearly a 50 year history in U.S. clinical laboratories.[1,2] During this time, its role as an educational, quality control, management or regulatory tool has been under continuous re-examination. For the many analytes requiring on-going PT participation and evaluation under CLIA'88, the question has been answered unequivocally: it is a regulatory tool.[3] However, even with CLIA'88 the strongest critics concede, and evidence today seems to support, PT still has an opportunity to fulfill educational, quality control and management roles. While it might be debated whether each role is enhanced or diminished by the regulatory focus, no one can deny that PT makes an impact on today's clinical laboratories.

Proficiency testing probably was best described by Forney as the "...distribution of (identical) unknown samples to laboratories for the purpose of determining the ability of laboratory personnel to achieve the correct analysis.[4]" Forney's definition incorporates the evaluation of accuracy through the interlaboratory assessment process. The most often used criterion for evaluation of accuracy is some form of the consensus "right answer." However, regardless of the criterion used to define "good" (acceptable) and "bad" (unacceptable) PT performance,

laboratories, by comparison to peers, have an opportunity to assess the quality (accuracy) of their performance.

When used in a regulatory context, the fundamental premise of PT is that if a laboratory performs acceptably in PT, it also analyzes patient samples correctly. The question that continues to plague the laboratory community, however, is related to the reliability of PT as an indicator of intralaboratory quality. Certainly at a minimum, PT is a means of assessing at least one form of accuracy. It is generally agreed that PT does not measure precision with any degree of reliability. However, the underlying question relates to the suggestion that PT samples are treated differently than routinely processed patient samples. Most recently, with the enactment of CLIA'88 and specified acceptable performance, the question of the validity of the evaluation criteria has further compounded the reliability issue.

PT as an Educational Tool

As originally envisioned by Belk and Sunderman, the PT process was educational, used to apprise laboratory directors as to when their analytical processes varied from that of the collective, wisdom of the group.[5] Belk and Sunderman maintained that competent laboratory directors would take appropriate corrective action when a problem was identified. Curiously under CLIA'88, PT failures mandate that the laboratory director develop a plan of correction. In 50 years, PT has not strayed too far from the original concept.

The College of American Pathologists (CAP) began offering a limited number of voluntary, interlaboratory (i.e., PT) surveys on an organized basis as early as 1947.[2] The obvious success of the PT process in improving laboratory quality led to a proliferation of voluntary interlaboratory testing programs by CAP, professional societies, and state and even municipal health departments during the 40's and 50's.[2,6] With the enactment of CLIA'67, PT became a mandated, but still primarily a self-directed, improvement process for large hospital and reference laboratories.[7,8] The rationale for mandating PT participation was that if a laboratory director could use data from PT as a means of self-assessment of quality, regulators could use the information for the same purpose. While CLIA'67 was relatively vague on what constituted acceptable levels of performance, and even left the PT providers to interpret data in terms of satisfactory or unsatisfactory performance, CLIA'88 does not. The step from a self-assessment to a minimum standard, performance requirement took place when the CLIA'88 regulations, as proposed by CDC and HCFA, included specific performance criteria.

PT as a Quality Control Tool

Intralaboratory error consists of two components - imprecision and inaccuracy. Laboratories assess imprecision through daily intra-laboratory QC activities, leaving the inaccuracy component to be assessed by some other means. For most laboratories accuracy is assessed through PT. PT, especially in Europe, also is called external QC. In the U. S., QC tends to focus more on standard deviations, or imprecision, and concentrates on achieving stable performance. The statistical mean determined in the QC process monitors drift and is not used to assess accuracy. Under CLIA'88, however, PT is linked to the broad area of QC. The clear implication is to make imprecision assessment and accuracy

monitoring a part of the QC process.

An offshoot of PT is related to peer comparison data. Regionalized QC programs such as those originally made popular by Hyland, Dade and General Diagnostics in the 1970's and, now CAP's QAS program, compare performance for laboratories analyzing the same lot of QC material.[9] Similar to PT, these programs, through mean and standard deviation comparisons, t-tests, Youdon plots, etc., offer an accuracy assessment in addition to a daily evaluation of imprecision.

PT as a Management Tool

Some of the ground-breaking differencesincorporated into CLIA'88 focus on PT as a management tool. Belk and Sunderman originally decided that PT should alert the director to potential problems within the laboratory; CLIA'88 requires the director to review PT data, document the review, and also to approve the remediation of any problems identified by the process. As a management tool, PT, originally and today, primarily provides data to determine: 1) the relationship of a given laboratory to peer laboratories, usually peer laboratories using the same methodology; 2) the robustness of methods (good versus poor quality) by assessing the amount of variation and pass rate among the peer group, and 3) the relationship between methods. The latter is the topic of Dr. Laskey; method comparisons are overlaid by the problem of matrix effects which have long plagued PT programs attempting to understand methodology differences.[10]

PT- The Regulatory Lynchpin of CLIA '88

CLIA'88 broke new ground regulating all laboratories, approximately 160,000. A large number of these (70-80,000) perform moderate and high complexity testing requiring PT participation. The drafters of the CLIA regulations should perhaps be applauded for the wisdom of their approach. As managers, they have devised a method for someone else to provide and grade the samples and then to send, in electronic form, the final results to HCFA. HCFA uses the data to accomplish its goal of assessing participant performance. This is not a statement of malevolent intent; it is a statement to acknowledge successful management practice, tempered only by the question as to whether PT should be used for regulation at all.

Critical Assessment of PT Performance Under CLIA'88

All laboratories, including physicians' office laboratories (POL), were required to enroll in a HCFA approved PT program by January 1994. Before 1994, PT participation for POLs was voluntary and results were used for educational purposes. PT data available from Wisconsin's HCFA-approved PT program and California's program indicate that significant performance problems exist, particularly for laboratories participating in PT for the first time.[11-13] In Wisconsin's program for example, 15% of all POL participants failed cholesterol on the initial survey in 1994 and 12% failed on the second survey. However, the good news was that only about 3 % of the laboratories failed both surveys. This indicates to us that most laboratories experiencing problems corrected them by the second survey. The data also show, however, that a large number of laboratories have rather marginal performance. The cholesterol performance limits are relatively generous, i.e., target value plus or minus 10%. Dr. Karen Nickel

reports similar findings in California, where 29% of all CLIA-certified POLs failed, that is, were unsuccessful in two out of three successive PT surveys, for at least one analyte.

Wisconsin's preliminary data from the second year of POL participation indicate continued improvement; further results from the California program are not yet available. Basing our assumption on these data, however, we would project that Belk and Sunderman will again be proved correct and performance will continue to improve. Interestingly, while PT data indicate a definite need for laboratory improvement and demonstrate that PT is an effective mechanism to achieve improvement, the wise men in Washington are seriously thinking of abandoning the process, at this point in time, for POLS. This is clearly brilliant thinking!

PT in the Year 2000 and Beyond

Objectively evaluating the PT process, one can see some positive attributes. Criticisms,however, include the fact that PT is expensive, time-consuming and disruptive to laboratory service. In addition, the number of samples are too small for meaningful interpretation, the process is flawed in that "good" laboratories sometime fail and "bad" laboratories may pass, the evaluation criteria may not be appropriate, and the PT sample matrix affects results. While the list of criticisms goes on, the lack of timeliness between analysis and result evaluation is, perhaps, the biggest drawback, preventing the achievement of PT's full potential as a quality assessment activity.

Making a quantum improvement in timeliness is, in our view, critical to the future of PT. As visionaries, we must not be afraid to dream. The information superhighway is in place. The possibility of reporting and evaluating PT results in "real time" opens the door to a whole new paradigm, one with tremendous opportunity for both laboratories and regulators. The PT process then could combine the attributes of both QC and PT into a single process, enhancing the cost effectiveness of the quality assurance activity. Appropriate computer algorithms, along with new designs of products, open the possibility of assessing multiple aspects of quality, all in real time and on-line, including accuracy, precision, linearity, reportable range, sensitivity, specificity, and method comparisons.

The vision for PT in the future should not be limited to what it can be; we should instead focus on what we want it to be. We must have the courage to achieve the dream.

References

1. Sunderman FW. The origin of proficiency testing for clinical laboratories in the United States. In *Proceedings of Second National Conference on Proficiency Testing.* Information Services, Bethesda, MD. 1975.

2. Dorsey DB. The evolution of proficiency testing in the USA. In *Proceedings of Second National Conference on Proficiency Testing.* Information Services, Bethesda, MD. 1975.

3. U.S. Department of Health and Human Services. Medicare, Medicaid and CLIA programs: Regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Final rule. Fed Regist. 1992; 57:7002-186.

4. Forney JE, et al. Laboratory evaluation

and certification. In *Quality Assurance Practices for Health Laboratories.* SL Inhom, ed., American Public Health Association, Washington, DC. 1977.

5. Belk WP, Sunderman FW. A survey of the accuracy of chemical analysis in clinical laboratories. *Am J Clin Pathol.* 1947; 17:853-861.

6. Skendzel LP, Copeland BE. An international laboratory survey. *Am J Clin Pathol. 1975* 63:1007-1011.

7. Department of Health, Education and Welfare, Public Health Service, Centers for Disease Control. Clinical Laboratory Improvement Act of 1967. Part F, Title 111, PHS §353.

8. Federal Health Insurance for the Aged: Regulations for Coverage of Service of Independent Laboratories. Public Health Report 13. Washington, DC: Department of Health, Education, and Welfare; 1968.

9. Lawson N, Rej R. The Validity of Proficiency Testing, and Performance in Proficiency Testing. *In Proceedings of Frontiers in Laboratory Practice Research,* Centers for Disease Control and Prevention, Atlanta, GA. 1995.

10. Laskey F. Achieving and assessing acceptable analytical performance: The challenge of matrix effects. In *Proceedings of Frontiers in Laboratory Practice Research,* Centers for Disease Control and Prevention, Atlanta, GA. 1995.

11. Burmeister BJ, Lanphear BT, Ehrmeyer SS, et al. Actual 1994 performance data from Wisconsin's HCFA-approved proficiency testing programs: Implications forlaboratories and regulations. *Clin Chem.* 1995;41:S210 [Abstract]

12. Ehrmeyer SS, Burmeister BJ, Laessig RH. Laboratory performance in a state proficiency testing program: what can a laboratorian take home? *Clin Immunoassay.* 1994;17:223-230.

13. Auxter S. POL exemption debate to begin on Capital Hill. *Clin Lab News.* 1995;21:1,12.

Summary of Workshop 1: Proficiency Testing

Facilitator: Paul Bachner, MD
Department of Pathology
University of Kentucky Hospital
Lexington, Kentucky

DLS Liaison: Shahram Shahangian, Ph.D.

Key Questions:

1) Does proficiency testing provide a reliable measure of actual laboratory performance?
2) How can the validity and utility of proficiency testing be enhanced?

Abstract: This workshop reviewed the benefits and drawbacks of current proficiency testing models and considered alternative approaches to proficiency testing. The latter included multi-program characterization of laboratory performance and performance evaluation based on retrospective reference method analyses of specimens previously tested by participating laboratories. Workshop participants also endorsed further study of "hybrid" quality assessment systems in which proficiency testing and quality control activities were blended. The participants concluded that proficiency testing is an important but incomplete measure of laboratory performance, and that multi-programmatic characterization of laboratory performance and restructuring of current proficiency testing models should be actively pursued.

The Workshop Session addressed two key thematic questions: (1) Does proficiency testing (PT) provide a reliable measure of actual laboratory performance? and (2) How can the validity and utility of PT be enhanced?

We began with a review of the seminal publication of Belk and Sunderman,[1] which described a survey performed 50 years ago and inaugurated the modern era of proficiency testing. Testing for common chemical and hematological analytes by 50 volunteer laboratories in Pennsylvania showed sufficient variation in results to place patients into very different clinical management scenarios. Subsequently, laboratory directors who were questioned about possible reasons for poor performance suggested, in order of frequency, poorly-trained and inadequate numbers of technicians, lack of understanding between pathologists and staff, poor institution floor space, and other miscellaneous reasons. Workshop participants observed that much progress has occurred in the intervening years and that the potential for degradation (and improvement) of performance is always present in the process components of laboratory practice.

Four invited presentations (published elsewhere in these Proceedings) developed the substrate for subsequent discussions. The first of these by Dr. Robert Rej of the New York State Department of Health (NYSDH) addressed the question of whether PT measures natural test performance. He stressed that the only components of the complete testing cycle that are

mirrored by PT are analysis and calculation, whereas non-analytic steps are reflected partially or not at all. Furthermore, process inconsistencies and analytic substrate (matrix) differences may contribute to a potential for enhanced or degraded performance. Dr. Rej also reported on a new initiative at the Wadsworth Center (NYSDH) dubbed "Retro-PT" in which hand-carried samples are analyzed by laboratories and subsequently re-analyzed by reference methods at the Wadsworth Center. These studies demonstrated relatively small biases with little potential for clinical impact and that the analyte-specific bias magnitude was similar to those noted in NYSDH PT testing programs.

Dr. Noel Lawson discussed multi-programmatic characterization of laboratory performance based on College of American Pathologists (CAP) data. He presented new data showing that relative PT performance is consistent over time (particularly poor performance) and across PT programs (linearity), performance is better in laboratories participating in interlaboratory PT programs and that similar biases are noted. His data also demonstrated that performance is better as a function of the length of time that a laboratory has been enrolled in PT testing and in laboratories that have been CAP inspected and accredited. These observations were noted in several programs and in multiple studies performed in 1984-86, 1988-90, and 1991-94.

Dr. Fred Lasky of Johnson & Johnson Clinical Diagnostics, Inc., spoke of performance problems associated with the manufacture of PT materials, specifically the impact of "matrix effects" detected in PT surveys but not noted with fresh patient samples. Dr. Lasky presented several recommendations and conclusions including that 1) fresh specimens are better for assessing performance, 2) manufacturing process improvement is possible, and 3) the relative benefit of such changes is questionable in view of the greater need to concentrate on improving non-analytic performance factors.

The final speaker, Dr. Sharon Ehrmeyer, of the University of Wisconsin, summarized the state-of-the-art and identified a "vision" for the future of PT not constrained by present programmatic limitations. Her presentation emphasized that current PT is expensive, time-consuming, disruptive, not timely, and provides incomplete identification of performance problems, even in "good" laboratories.

Based on these presentations, the lively and extensive discussions of participants resulted in several broad areas of agreement that are summarized as follows:

- A NEED TO RESTRUCTURE PT TO BETTER ASSESS NON-ANALYTIC PERFORMANCE: Although recognizing the difficulty of achieving this recommendation, the need to move toward this goal will be accentuated by the increasing utilization of point-of-care testing and the decentralization of testing. The potential for rapid, real-time data transfer through the information highway will encourage emerging regional and local initiatives that try to blend PT and quality control efforts within integrated health care delivery systems.

- WE ARE ON THE THRESHOLD OF HYBRID PT AND QUALITY CONTROL (QC) CONTROL SYSTEMS and that it will be important to validate and compare performance

between different models of what workshop participants provisionally identified as "inter-community PT/QC." To support this trend it will be necessary to look beyond current PT models as well as explore "alternative" QC practice that will be more appropriate to emerging point-of-care technology and instrumentation.

- A CONTINUING ROLE FOR CURRENT NATIONAL PT to provide traceability for "hybrid" PT/QC in local/regional as well as evolving national networks based in large hospital systems and commercial laboratories.

- THE CONTINUING BENEFIT OF CURRENT PT to provide a useful estimate of the state-of-the-art, a system to monitor individual laboratory quality improvement and an early-warning system for problem identification, to satisfy the need for independent assessment, as well as to meet regulatory requirements and considerations of public accountability.

- THE VALUE OF MULTI-PROGRAM CHARACTERIZATION as presented by Dr. Lawson was endorsed and it was stated that hybrid PT/QC could be viewed as a variant of multi-program characterization of laboratory performance that could incorporate blind and split sample testing as components.

Additional important observations made by Institute participants during the presentation of the workshop proceedings to the Plenary session were the serious potential for compromise of the educational and quality improvement role of PT by continuing emphasis on the regulatory and punitive aspect of PT. Such an emphasis will divert PT program providers from designing challenges that will encourage laboratory improvement in order to avoid excessively high failure rates. A related observation was the need for PT program providers to recognize the participating laboratory as their prime customer, rather than government regulatory agencies.

In summary, the Workshop participants concluded, in response to the key questions that were posed, (1) that PT is an important indicator of laboratory performance, but only one of other (QC, quality assurance, inspection, personnel standards, patient test management) partial and incomplete measures of laboratory quality, and (2) that the validity and utility of PT should be enhanced by an increased emphasis on multi-programmatic characterization of laboratory performance and by the restructuring of PT (and hybrid PT/QC programs) to better reflect the
entire testing cycle.

References

1. Belk WP, Sunderman FW. A survey of the accuracy of chemical analyses in clinical laboratories. Am J Clin Pathol 17:853-861, 1947.