Clinical Chemistry 57:12
1670–1680 (2011)

Reviews

Proficiency Testing/External Quality Assessment: Current Challenges and Future Directions

W. Greg Miller,[1*] Graham R.D. Jones,[2] Gary L. Horowitz,[3] and Cas Weykamp[4]

BACKGROUND: Proficiency testing (PT), or external quality assessment (EQA), is intended to verify on a recurring basis that laboratory results conform to expectations for the quality required for patient care.

CONTENT: Key factors for interpreting PT/EQA results are knowledge of the commutability of the samples used and the process used for target value assignment. A commutable PT/EQA sample demonstrates the same numeric relationship between different measurement procedures as that expected for patients' samples. Noncommutable PT/EQA samples frequently have a matrix-related bias of unknown magnitude that limits interpretation of results. PT/EQA results for commutable samples can be used to assess accuracy against a reference measurement procedure or a designated comparison method. In addition, the agreement of the results between different measurement procedures for commutable samples reflects that which would be seen for patients' samples. PT/EQA results for noncommutable samples must be compared to a peer group mean/median of results from participants who use measurement procedures that are expected to have the same or very similar matrix-related bias. Peer group evaluation is used to asses whether a laboratory is using a measurement procedure in conformance to the manufacturer's specifications and/or in conformance to other laboratories using the same technology. A noncommutable PT/EQA sample does not give meaningful information about the relationship of results for patients' samples between different measurement procedures.

SUMMARY: PT/EQA provides substantial value to the practice of laboratory medicine by assessing the performance of individual laboratories and, when commutable samples are used, the status of standardization or harmonization among different measurement procedures.

Proficiency testing (PT),[5] also called external quality assessment (EQA), was introduced into laboratory medicine more than 60 years ago *(1, 2)* as an educational tool to address observations that results for aliquots of the same sample were different when measured by different laboratories. The measurement procedures used at that time were laboratory developed and differed among laboratories in implementation and calibration details. PT/EQA results were used to stimulate standardization of procedures and calibrators to achieve more uniform results among laboratories. PT/EQA programs have evolved in scope and sophistication and are now an essential component of a laboratory's quality management system. PT/EQA is intended to verify on a recurring basis that laboratory results conform to expectations for the quality required for patient care. PT/EQA is a component of laboratory accreditation requirements in many countries.

The spectrum of PT/EQA includes analytical performance and pre- and postanalytical components *(3)*. An international standard has been published that provides management information and requirements for PT/EQA providers on organizing and conducting such programs *(4)*. In this review we focus on key issues in the design, performance, and interpretation of PT/EQA schemes with the aim of describing strengths and limitations of current schemes and explaining how PT/EQA can contribute to improvements in laboratory medicine. This review is limited to PT/EQA for assessment of quantitative measurement procedures.

In general, a PT/EQA survey is conducted by sending a set of samples from an organizing body to a group of participating laboratories for measurement of 1 or more analytes present in the samples. The PT/EQA samples are intended to simulate the clinical samples usually measured. Laboratories are not informed of the analyte concentration or activity in a particular sample

[1] Virginia Commonwealth University, Richmond, VA; [2] St Vincent's Hospital and University of New South Wales, Sydney, Australia; [3] Harvard Medical School, Boston, MA; [4] Queen Beatrix Hospital, Winterswijk, the Netherlands.
* Address correspondence to this author at: P.O. Box 980286; Richmond, VA, 23298-0286. Fax 804-828-0375; e-mail gmiller@vcu.edu.
Received May 12, 2011; accepted August 8, 2011.
Previously published online at DOI: 10.1373/clinchem.2011.168641

[5] Nonstandard abbreviations: PT, proficiency testing; EQA, external quality assessment; IVD, in vitro diagnostic.

and perform measurements in the same manner as for patient samples. Results for the samples are returned to the PT/EQA organizer for evaluation of conformance to the expected results. The organizer prepares a report that includes the results reported by a laboratory, the method used for the measurements, the target values expected for each analyte, and an evaluation of whether the individual laboratory's results met the performance requirements. Reports may also include evaluation of the performance of the various measurement procedures used by the participants.

Samples Used for PT/EQA

Ideal samples for a PT/EQA program would fulfill a range of criteria: stable for the conditions under which they will be transported and stored, homogeneous across all the aliquots produced, have analyte concentrations that include the expected clinical range, include appropriate sample types (e.g., urine, whole blood, serum), available in sufficient volume, inexpensive enough for cost not to be an impediment, and behave in clinical laboratory measurement procedures in the same manner as patient samples. In practice, it is impossible to achieve all these goals, and some compromises are required in the preparation of PT/EQA materials. Commutability with clinical patient samples is one of the most important concepts affecting the design and interpretation of PT/EQA schemes.

COMMUTABILITY

Commutability is a property of a PT/EQA sample whereby the sample has the same numeric relationship between measurement procedures as is observed for a panel of representative clinical patient samples *(5–8)*. The concept of commutability is illustrated in Fig. 1A, which shows the relationship between 2 measurement procedures for a panel of individual patient samples. In this example, the numeric relationship is established by regression analysis, and the 95% prediction interval defines the expected statistical distribution for results from commutable PT/EQA samples that have the same numeric relationship as do the patient samples *(7)*. Fig. 1B shows that results for noncommutable PT/EQA samples are outside the prediction interval.

A PT/EQA sample that is commutable gives a numeric result that is equivalent to that expected for a patient sample containing the same quantity of an analyte among different measurement procedures. A PT/EQA sample that is not commutable for different measurement procedures does not give meaningful information about the relationship of results for a patient sample among those procedures. Numerous investigations have reported that approximately half the samples examined have not been commutable with

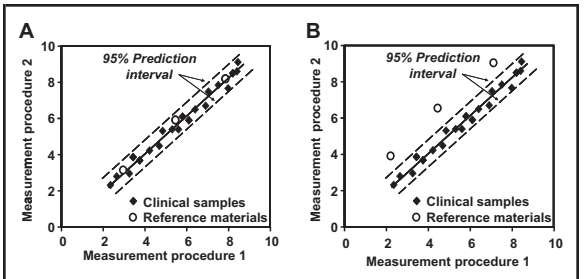

Fig. 1. Example of the commutability property of PT/EQA samples.

(A), Typical results for commutable PT/EQA samples that have results within the prediction interval based on a panel of individual clinical patient samples. (B), Results for noncommutable PT/EQA samples that are outside the prediction interval based on a panel of patient samples.

clinical patient samples *(5, 6, 9–13)*. The terms "matrix-related bias" and "matrix effect" are used to refer to the component of bias that is caused by noncommutability.

In PT/EQA, the terms noncommutability, matrix-related bias, and matrix effect are used to refer to differences that occur only in the PT/EQA samples but not in authentic clinical patient samples. Consequently, in PT/EQA, interference from an endogenous substance present in abnormal concentrations (e.g., high bilirubin) is generally not considered a matrix effect. However, noncommutability caused by a nonnative form of an analyte (e.g., ditauro bilirubin) is considered a matrix effect.

PREPARATION OF SAMPLES INTENDED TO BE COMMUTABLE

Samples intended to be commutable are typically prepared by collection and processing of the intended sample type in the same manner as for clinical samples, followed by division into aliquots and distribution under stable conditions. Samples from a single donor or pooled samples from multiple donors can be used. The number of aliquots needed and the desired concentrations or activities of analytes frequently preclude single-donor samples. Single-donor samples have the limitation that an interfering substance may be present that influences 1 or more of the measurement procedures, thus confounding interpretation. Pooled samples will dilute an interfering substance and, depending on how many donor samples contribute to the pool, may eliminate its influence. However, pooled samples have a potential limitation that interactions of components such as serum proteins or urine complexes from different donors may cause aggregation or precipitation that necessitates further processing and potential modification of the matrix.

The procedures for collection and handling of samples are critical to avoid influencing the matrix and to preserve commutability in the final aliquots. The CLSI guideline C37A describes a rigorous protocol to collect blood, obtain serum, prepare a pool, and freeze aliquots under conditions that do not alter the commutability characteristics for cholesterol *(14)*. This protocol has also been validated to be suitable for triglycerides and HDL cholesterol *(15)* and for creatinine *(16)*. The C37A protocol has not been validated to produce samples that are commutable with patient samples for other analytes but represents the best available approach and has been used to prepare single-donor serum samples and pools for several investigations of the trueness of measurement procedures for several analytes *(17–21)*. Thienpont et al. found that processing of serum (e.g., sterile filtration, storage before aliquoting and freezing) may disturb the equilibrium between free and protein-bound thyroid hormone and hence jeopardize the commutability of a reference material prepared even from sera that has been subjected to minimal processing *(22)*. Rigorous protocols for other matrices have not been reported, but the general principle to collect unaltered samples, pool them, and either distribute and measure aliquots immediately or freeze aliquots at ≤ -70 °C is accepted as the best available approach to obtain PT/EQA samples likely to be commutable.

A limitation of donor samples is that desirable concentrations or activities of analytes may not be available. Higher concentrations or activities can be achieved by adding analytes to pooled unaltered samples. It can be hypothesized that supplementation with purified analytes does not alter the matrix and the samples will remain commutable. This assumption has been reported to be true for creatinine added to a serum pool *(16)*. However, it is a reasonable assumption for simple analytes. Confidence in that assumption declines as the analyte becomes more complex or may not be available in highly purified form, or if the matrix of the supplement contributes to alteration of the matrix of the native samples. Lower concentrations of an analyte may be achieved by removing an analyte, for example by immunoabsorption on a solid phase. Analyte removal procedures may remove unintended molecules or otherwise modify the matrix, especially when nonspecific techniques such as charcoal or protein-A are used.

VALIDATING COMMUTABILITY OF SAMPLES

There are consensus procedures to validate commutability of reference materials that are applicable to PT/EQA samples *(6, 7)*. It is preferable to validate the commutability of PT/EQA samples with unaltered single-donor samples for all combinations of measurement procedures for which that sample is intended to be used. However, it can be difficult to obtain the clinical samples and expensive to perform a commutability validation. For practical reasons, commutability of only 1 lot of a PT/EQA material may be evaluated, and subsequent lots prepared the same way may be assumed to be commutable.

In current practice, samples are commonly assumed to be commutable only on the basis of the stringency of their preparation, as described above. In such cases, the assumption is reasonable but the possibility of noncommutability remains a limitation in data interpretation. The assumption of commutability becomes less likely the more donor sample handling deviates from that used for typical clinical samples.

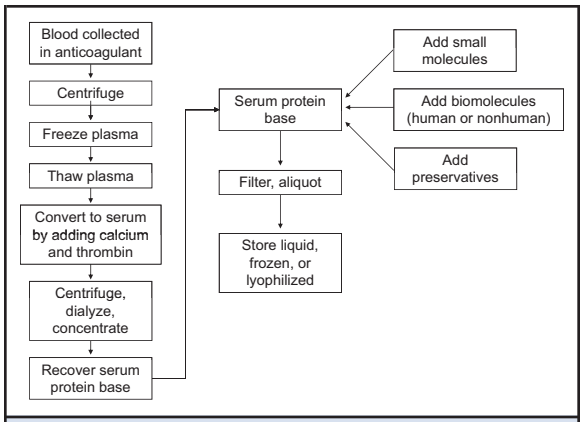

Fig. 2. Representative preparation steps for a serum-based PT/EQA sample to illustrate potential influences on its commutability.

PREPARATION OF SAMPLES UNLIKELY TO BE COMMUTABLE

During their preparation PT/EQA samples are frequently modified, which causes them to be noncommutable. Many different and frequently proprietary procedures are used by manufacturers to obtain PT/EQA samples with suitable analyte quantities and stability characteristics for storage and distribution. A representative preparation protocol for what is frequently termed "serum" in PT/EQA programs is shown in Fig. 2 as an example of some important potential influences on the matrix that may be introduced in manufacturing and that can affect the commutability characteristics of a sample. Noncommutability has been attributed to alteration of the sample matrix even if the sample originated or was derived from human sources, to nonnative forms of an analyte that produce a measurement signal different from that expected for native forms, to impurities introduced with analyte supplements, to preservation processes, and to other

influences not present in native clinical samples *(5, 6, 9, 23–25)*.

Target Values and Acceptance Criteria for PT/EQA Results

To interpret a PT/EQA result, the program organizer must provide a target value and a range for acceptable values around that target. It is important for users of such programs to be aware of the different techniques that may be used for these processes and their strengths and limitations.

ASSIGNMENT OF TARGET VALUES WHEN THE SAMPLES ARE COMMUTABLE

A key benefit with a commutable PT/EQA sample is assessment of traceability of the result to a reference system. For this purpose the value assignment for the PT/EQA sample must be made by using a reference measurement procedure or a high-specificity comparative method that is traceable to a reference measurement procedure. When available, traceability should be to methods, materials, and laboratories listed on the Joint Committee for Traceability in Laboratory Medicine database *(26)*. Target assignment by value transfer based on results from certified reference materials is possible if the commutability of the reference materials has been verified *(8, 27–29)*. Assignment of targets on the basis of known amounts of a weighed-in material is dependent on the purity of the material, the accuracy of any measurement devices, and the demonstrated similarity between the pure material and the form of the analyte in human samples. Validation that the final samples including additions remain commutable with clinical samples is recommended. When a reference system is not available, the all-participant mean or median value after outlier exclusion may be used as the target because the same results are expected for all procedures for a commutable sample.

ASSIGNMENT OF TARGET VALUES WHEN THE COMMUTABILITY OF SAMPLES IS NOT LIKELY

The most common procedure used to assign a target value is to categorize participant methods into "peer groups" that represent similar technology and calculate the mean or median of the peer group as the target value after removal of outlier values. A peer group consists of methods that are likely to have the same matrix-related bias for a given PT/EQA sample and thus are expected to have the same results for that PT/EQA sample. It is common for peer groups to be formed as instrument/reagent groupings from the same manufacturer. A limitation when calculating the mean or median is the number of results in the peer group. As the number of results decreases, or the dispersion among the results increases, the uncertainty in the target value increases. For small peer groups, provided there is limited dispersion among results, comparison with the median can provide a useful assessment.

When the commutability of samples is unknown, a reference measurement procedure value is less useful because it is not possible to determine if a difference from that target value is caused by a calibration bias or by a matrix-related bias of unknown magnitude. A reference measurement procedure target is more likely to be useful for evaluation of measurement procedures with high analytical specificity and less useful for methods with poor analytical specificity. Similarly an all-methods mean/median is less useful as a target value unless there is evidence to support that all methods have a similar matrix-related bias or there is no other alternative (e.g., all method groups are very small). If an all-methods mean/median is used, the larger peer groups will have greater influence on the apparent target value, the value may change over time depending on the relative number of participants using different measurement procedures, and the target value is likely to be inappropriate for at least some of the peer groups being evaluated.

It may appear that a reference measurement procedure value or an all-methods mean/median is satisfactory as a target value because results for participants in different peer groups fortuitously meet the acceptance criteria. However, there is no scientific rigor behind such an approach. If a participant's sample results were to "fail" a comparison to a reference measurement procedure or an all-method mean/median target value, a legitimate explanation may be that the magnitude of a matrix-related bias for that peer group was different from the matrix-related biases for other peer groups. Failure of a measurement result to agree with either of these target values does not provide conclusive evidence that the results for clinical samples are not acceptable. However, apparent differences may or may not be attributable to matrix-related biases, and further investigation is required to determine if the differences are also observed for patient sample results.

ACCEPTANCE CRITERIA FOR PT/EQA RESULTS

Limits or quality standards around the target value are established against which performance can be assessed. In general, PT/EQA scheme limits may be considered regulatory, statistical, or clinical. Regulatory limits such as those required by the USA Clinical Laboratory Improvement Amendment *(30)* or the German RiliBAeK *(31)* tend to be wider, with the intent to identify laboratories with sufficiently poor performance that they should not be able to continue to practice. Statis-

tical limits (e.g., ±2–3 SD of the distribution of a participants' results) are based on the unstated assumption that the measurement procedures are suitable for clinical use and that performance is acceptable if it is concordant with others obtained by use of the same procedures. Clinically based criteria, for example based on a difference that may affect clinical decisions or on biological variation *(32, 33)*, are desirable but have proved difficult to implement *(34)*. Consequently, the criteria derived are highly variable among different schemes *(35)*.

There are sources of variability in PT/EQA results not found in patient results that may cause the acceptance limits to be larger than what is needed on the basis of clinical requirements. Differences in calibrator and reagent lot-to-lot consistency and differences in operation and maintenance of measurement procedures will contribute to between-laboratory variability. Degradation of samples during transportation and storage before measurement can affect results obtained for these samples differently than for samples collected in the clinical setting. For noncommutable samples, the magnitude of a matrix-related bias can be different for different lots of reagent causing a larger dispersion in results than would be observed for clinical samples *(36)*.

It is common for PT/EQA schemes to have a single set of limits for each analyte. The limits are used to assess individual results and must be considered total error limits because bias, imprecision, and analytical nonspecificity can contribute to the variation in a single result. It may be appropriate to have different limits to separately assess bias and imprecision when replicate samples are included. It is also important to recognize that most PT/EQA limits are set as a minimum standard to identify results that indicate poor performance. Thus, meeting these standards may not indicate that performance is optimal nor that performance meets all clinical needs. A separate set of limits may be required for assessment of whether results meet clinical needs.

Statistical limits (e.g., ±2–3 SD) compensate for some of the limitations in PT/EQA samples and create acceptance limits that have a predictable number of nonconformities. When fixed criteria are used, the uncertainty in the target value will be a fraction of the acceptance interval and may be an important consideration when the criteria are closely aligned with clinical requirements. When the acceptable interval is expressed as a percent, it may also be necessary to include a fixed unit interval below a concentration at which a percent is not reasonably achievable because the SD of a measurement procedure becomes a larger fraction of the acceptable interval. For example, acceptance criteria for alanine aminotransferase might be ±15% or ±6 U/L below 40 U/L; thus, a sample with a concentration of 20 U/L would then have an acceptable interval of ±6 U/L.

Participant Evaluation of Their Own PT/EQA Results

MEASUREMENT AND REPORTING OF PT/EQA SAMPLE RESULTS

In general PT/EQA schemes request that samples be analyzed as though they were patient samples. For example, samples should be run at varying times of the day by different technologists. Care must be taken to ensure appropriate sample handling so that analyte degradation does not contribute to the dispersion of results. There should be no attempt to produce "best" results by replicate analysis or testing immediately following internal QC or recalibration. Such practices compromise the primary objective of the PT/EQA process to evaluate the laboratory's performance for routine patient samples.

To ensure consistent results for patients whose samples may be measured by more than one measurement procedure in the same healthcare setting, a laboratory may have adjusted the calibration of one procedure to agree with another one. If commutable samples are used, the PT/EQA results can be reported because the calibration adjustment was intended to produce correct results for patient samples. However, if samples with unknown commutability are used, the PT/EQA results should be transformed to the manufacturer's intended calibration conditions to allow evaluation against the appropriate peer group target value. Such a transformation can be performed by measuring the PT/EQA sample as a patient sample and then back-calculating to remove the calibration adjustment factor to obtain a result that would have been produced by the method with the use of its original manufacturer's calibration condition.

INTERPRETING PT/EQA RESULTS FOR COMMUTABLE SAMPLES (ACCURACY-BASED EVALUATION)

Commutable PT/EQA samples have the desirable attribute that relationships among results will correspond to the relationships observed for clinical patient samples. Consequently, a laboratory can directly determine the accuracy of patient results by comparing PT/EQA results to those from a reference measurement procedure or from a designated comparison method. This arrangement is now referred to as accuracy-based evaluation. In contrast to noncommutable samples, commutable samples allow a laboratory to assess agreement with other measurement procedures and imprecision among all methods as well as within a method group that reflects the condition for patient samples. Commutable samples included in multiple survey events

give a reliable estimate of imprecision over time within a laboratory.

INTERPRETING PT/EQA RESULTS FOR SAMPLES WITH UNKNOWN COMMUTABILITY

Because of the noncommutability limitation, PT/EQA results are compared to the peer group mean/median of results from participants who use measurement procedures that are expected to have the same or very similar matrix-related bias. Supplemental Fig. S1 (in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol57/issue12) shows an example of a participant report. Peer group evaluation does not permit direct verification of the accuracy of a result to a reference measurement procedure, to a designated comparison method, or to an all-participant (or all-method) mean/median. Nonetheless, peer group evaluation provides valuable information to assess quality, verifying that a laboratory is using a measurement procedure in conformance to the manufacturer's specifications and to other laboratories using the same technology. In this situation, traceability of the measurement procedure's calibration to the highest order reference system is provided by the manufacturer. Consequently, verification that the PT/EQA results meet the manufacturer's specifications indirectly verifies the accuracy of patient results if it is assumed that the manufacturer has correctly calibrated an assay. However, agreement with a peer group may not detect an error by a manufacturer when all supplied calibrators in a region are affected *(37, 38)*. The SD in peer group evaluation allows participants to assess the effectiveness of a manufacturer's quality system to deliver uniform results among the users of that technology. Including the same sample in multiple survey events allows an estimate of the imprecision over time within a laboratory.

Statistically based criteria have the undesirable property that the acceptance limits may vary between peer groups measuring the same analyte. Imprecise-method peer groups will have a large interval for acceptable results and there is little incentive for participants to change to better methods. A very precise-method peer group will have a small interval for acceptable results that may be smaller than is required for clinical needs, and some participants will not meet the criteria although the results are acceptable for clinical care. On the other hand, some analytes have diagnostically important physiologic changes in concentration that are smaller than the SD of most clinical measurement procedures. Fig. 3 illustrates these limitations for creatine kinase and calcium as they relate to acceptance criteria.

PARTICIPANT FOLLOW-UP ON PT/EQA RESULTS

Each unacceptable PT/EQA result must be investigated and the findings and any corrective action documented. It is recommended that a laboratory follow up on results that may have been within the acceptance criteria, but were statistically less probable to be correct and may indicate an impending error condition. For example, results that are >2.5 SD from a target value may be acceptable in a survey but may still indicate a possible problem that should be investigated. In addition, the results for all PT/EQA results for an analyte in a set should be considered in the investigation because multiple results with relatively large differences scattered on both sides of the target values suggest inadequate precision, whereas multiple results with relatively large differences in the same direction from the target values suggest bias. Other analytes should also be reviewed because this may lead to the identification of deficiencies in sample handling or preparation.

It is recommended that trends in results for PT/EQA samples from different survey events be monitored over time. Many programs provide a graphical representation of results over a time interval, which assist in identification of a systematic bias although individual results may not have been scored as unacceptable (see online Supplemental Fig. S1 for an example). In such a case, the laboratory can initiate corrective action before an impending problem becomes a problem of clinical significance.

Table 1 provides a classification of the types of problems that may be identified by PT/EQA results on the basis of a consensus guideline from the CLSI on PT/EQA for the clinical laboratory *(3)*. After exclusion of clerical errors, steps in the investigation typically include:

- Gather data related to the testing event to include records of calibration, reagent use, QC results, and maintenance procedures;
- Obtain other data on assay performance, e.g., previous PT/EQA results and relevant patient data;
- Identify the root cause of the error;
- Take corrective action and preventive action if indicated;
- Monitor the success of the corrective action;
- Document the investigation and the corrective action.

It is important to recognize that a PT/EQA result represents 1 point in time and will occasionally be a random error. It is common practice to repeat the measurement using a stored aliquot of the PT/EQA sample that had an unacceptable result (assuming the measurand was stable on storage) as well as the other PT/EQA samples from the same set to confirm if the problem has persisted or to conclude that the problem no longer exists and the original unacceptable result was a

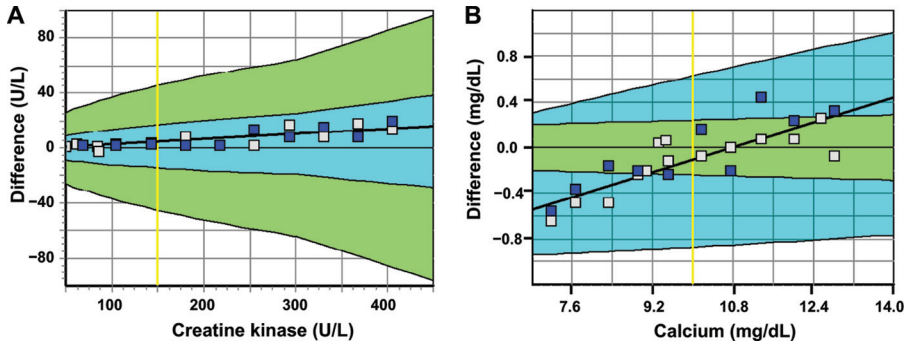

Fig. 3. Example of an annual PT/EQA report for creatine kinase (CK) (A) and calcium (Ca) (B).

This scheme measures 12 samples at 2-week intervals with duplicate samples used in the second half-year. Samples were prepared by pooling residual sera from the routine clinical chemistry laboratory (criteria: nonicteric, nonlipemic, <72 h at 4–8 °C) and storing at −84 °C. The pools were thawed and calcium chloride and recombinant human CK were added to create higher quantities. Then the samples were mixed, dispensed, and frozen at −84 °C within 1 working day. Commutability was verified according to the twin-study approach [Baadenhuijsen H et al. *(44)*]. Samples were shipped on dry ice and stored by the participants below −70 °C until used. The format presents the performance over a 1 year period based on acceptance criteria for both clinical total allowable error (green zone) and for state-of-the-art based on the distribution of 90% of the results of all laboratories (blue zone). On the *x* axis is the target value established with a reference measurement procedure. On the *y* axis is the deviation from the target. The vertical yellow line represents a clinically relevant decision concentration. Open white squares represent results from the first half-year and filled blue squares represent results from the duplicate samples in the second half-year. For CK with a large biological variation, the clinical acceptance limits are wider than the distribution based limits. For Ca with a small biological variation, the clinical limits are narrower than the distribution-based limits. Adapted with permission from the general chemistry EQA program of the SKML, the EQA organizer in the Netherlands.

random event. A single random error is not conclusive, may not persist, and therefore no corrective action is indicated. If the repeated result is still unacceptable, the laboratory concludes that a systematic error is present, conducts further investigation to identify the root cause, and then initiates corrective action.

Using PT/EQA to Assess Measurement Procedure Performance

PT/EQA WITH THE USE OF COMMUTABLE SAMPLES (ACCURACY-BASED EVALUATION)

PT/EQA programs that use commutable samples are particularly valuable for in vitro diagnostic (IVD) manufacturers, for laboratories that develop their own measurement procedures, and for standardization/harmonization programs. PT/EQA results can be used to evaluate the success of calibration standardization to a reference measurement procedure or calibration harmonization when no reference measurement procedure is available *(9–13, 39, 40)*. In addition, information is provided on the effectiveness of a manufacturer's transfer of calibration traceability to routine measurement procedures in the field.

Results from commutable PT/EQA samples reflect relationships expected for patient samples because there is no significant matrix-related bias. The mean/median values for different methods can be compared to each other and to results from a reference measurement procedure, a designated comparison method, or an all-participant mean/median to assess the uniformity of results for patient samples among different measurement procedures. A procedure with aberrant results can be identified, and the manufacturer can correct the calibration to conform to the appropriate standard. The SD for a method will be influenced by the same factors that influence imprecision for patient samples and, consequently, gives the best available information on the effectiveness of a manufacturer's quality system to deliver uniform results among different users and informs users which technologies have better precision and uniformity among laboratories.

PT/EQA data can be used to inform professional bodies on decisions regarding use of pathology results. Recent examples of analyses for which laboratory quality has been assessed, improved, and then reassessed to reach appropriate clinical requirements include serum creatinine for calculation of the estimated glomerular filtration rate *(41)* and hemoglobin A_{1c} assays for diagnosis and monitoring of diabetes *(42)*. An example of the role of PT/EQA to improve performance of hemoglobin A_{1c} measurement procedures is described in online Fig. S2.

Table 1. Classification of potential problems identified when investigating unacceptable PT/EQA results.[a]

1. Clerical errors
Incorrectly transcribed PT result from the instrument read-out to the report form
PT sample was mislabeled in the laboratory
Incorrect instrument or method was reported on the results submission form
Incorrect units were reported
Decimal point was misplaced
2. Methodological problems
Inadequate standard operating procedure
Problem with manufacture or preparation of reagents or calibrators (e.g., unstable)
Lot-to-lot variation in reagents or calibrators
Incorrect value assignment of calibrators
Method lacks adequate specificity for the measurand
Method lacks adequate sensitivity to measure the concentration
Carryover from a previous sample
Inadequate QC procedures used
3. Equipment problems
Obstruction of instrument tubing/orifice by clot
Misalignment of instrument probes
Incorrect instrument data processing functions
Incorrect instrument settings
Automatic pipettor not calibrated to acceptable precision and accuracy
Equipment component malfunction, e.g., light source, membrane, fluidics, detector
Incorrect instrument conditions, e.g., water quality, surrounding temperature
Instrument maintenance not performed appropriately
4. Technical problems caused by personnel errors
Did not operate equipment correctly or did not conform to method standard operating procedure
Incorrect storage, preparation or handling of reagents or calibrators
Delay causing evaporation or deterioration of the PT sample
Failure to follow recommended instrument function checks or maintenance
Pipetting or dilution error
Calculation error
Misinterpretation of test result
5. A problem with the PT material such as
Incorrect storage, preparation, or handling of PT materials
Differences between PT samples and patient samples, e.g., matrix, additives, stabilizers
Sample deteriorated in transit or during laboratory storage
Sample had weak or borderline reaction
Sample contained interfering factors (which may be method specific)
Sample was not homogeneous among vials

[a] This classification scheme assists in the development of an appropriate corrective action plan. Adapted, with permission, from CLSI *(3)* and from Miller *(45)*.

PT/EQA providers should be encouraged to publish such information because it can be used to confirm the validity of combining data as we move to greater sharing of patient results via electronic and other means.

PT/EQA PERFORMED BY USING SAMPLES WITH UNKNOWN COMMUTABILITY

The quantitative relationships between the mean/median values for different measurement procedures, including the relationship to a reference measurement procedure, cannot be determined from PT/EQA results for samples likely to be noncommutable owing to the approximate 50% frequency of observed matrix-related biases *(5, 6, 9–13)* and the unknown magnitude of these biases. An example of erroneous conclusions regarding vitamin D measurements based on noncommutable PT/EQA results is shown in Table 2. The apparent differences among the peer groups for the conventional noncommutable sample are artifacts of different magnitudes of matrix-related biases because the peer groups have nearly the same values for the commutable sample. In situations in which a matrix-related bias is quantitated for a given sample and measurement procedure, it may be possible to evaluate performance by using a correction factor for the matrix-related component of bias *(39, 43)*.

FUTURE DEVELOPMENT OF PT/EQA PROGRAMS

PT/EQA programs can be classified into 6 categories according to how well they are able to evaluate performance (Table 3). Evaluation capability depends on 3 characteristics: sample commutability, process for target value assignment, and inclusion or noninclusion of replicate samples. Category 1 is the most desirable because programs in this category use commutable samples with target values established by a reference system and can evaluate both individual laboratories and measurement procedures for reproducibility, calibration traceability, and uniformity between laboratories and between measurement procedures. Programs in category 2 have the same attributes as category 1 except that within-laboratory reproducibility cannot be evaluated because replicate samples are not used within a survey cycle. Programs in categories 3 and 4 also use commutable samples but, because the target values are not established by a reference system, the evaluation is limited to the uniformity among results (harmonization), a feature of considerable value for laboratory medicine. Programs in categories 5 and 6 use samples likely to be noncommutable, thereby limiting evaluation to peer-group comparisons and failing to provide information on bias between different measurement procedures.

Table 2. Results for 25-hydroxyvitamin D for noncommutable and commutable PT/EQA samples.[a]

	Conventional noncommutable PT/EQA sample			Commutable fresh-frozen serum sample		
Peer group	Participants, n	Mean, ng/mL	CV	Participants, n	Mean, ng/mL	CV
1	25	119.8	58.2%	8	23.5	12.3%
2	108	97.6	11.65%	53	25.9	10.5%
3	19	51.2	15.3%	12	30.1	12.9%
4	24	55.9	19.8%	15	26.4	23.6%

[a] Abstracted with permission from Survey 2009 Y-A of the College of American Pathologists. The data are also available online at http://www.cap.org/apps/docs/committees/chemistry/measurements_25_OH_vitamin_d.pdf (accessed July 4, 2011). All samples were determined by mass spectrometry to contain 100% 25-hydroxyvitamin D_3, so differences between peer groups cannot be ascribed to different sensitivities to 25-hydroxyvitamin D_2 vs D_3.

Ideally all PT/EQA programs would be category 1 schemes. Unfortunately, however, category 1 programs are rare because of constraints including:

- Technical aspects such as a lack of reference measurement procedures, absence of certified reference materials, inability to prepare commutable samples;
- Practical considerations such as the difficulty of preparing samples covering the full measuring interval and the complicated logistics of preparation and distribution of fresh or frozen samples;
- Psychological limitations such as lack of awareness of the quality factors important in PT/EQA or unwillingness to adopt these;
- Economic concerns because distributing commutable samples in sufficient quantity and providing target values with reference measurement procedures is expensive.

The challenge for PT/EQA organizers is to overcome these limitations. The responsibility for the quality of laboratory testing is now a shared responsibility between the individual laboratory, the IVD industry, reference laboratories, and professional organizations. Consequently, the goals for an optimal PT/EQA program are to evaluate bias and reproducibility throughout the measuring interval for an individual laboratory, and calibration traceability and uniformity between laboratories for the measurement procedures used.

The global adoption of clinical practice guidelines requires those of us in the clinical laboratory profession to produce, and to verify that we are producing, glob-

Table 3. Evaluation capabilities of PT/EQA related to scheme design.

				Evaluation capability						
				Accuracy						
				Individual laboratory					Standardization or harmonization[b]	
	Sample characteristics				Relative to participant results		Reproducibility		Measurement procedure calibration traceability	
Category	Commutable	Value assigned with RMP[a] or CRM	Replicate samples in survey	Absolute vs RMP or CRM	Overall	Peer group	Individual laboratory intralab CV	Measurement procedure interlab CV	Absolute vs RMP or CRM	Relative to participant results
1	Yes	Yes	Yes	X	X	X	X	X	X	X
2	Yes	Yes	No	X	X	X		X	X	X
3	Yes	No	Yes		X	X	X	X		X
4	Yes	No	No		X	X		X		X
5	No	No	Yes			X	X	X		
6	No	No	No			X		X		

[a] RMP, reference measurement procedure; CRM, certified reference material.
[b] Standardization when patient results are equivalent between measurement procedures and calibration is traceable to SI by use of a reference measurement procedure; harmonization when patient results are equivalent between measurement procedures and calibration is not traceable to a reference measurement procedure.

ally equivalent results. Initiatives to standardize or harmonize measurement procedure results *(40, 44–45)* require surveillance by PT/EQA schemes that use commutable samples and target values set according to reference measurement procedures when available or by use of consensus approaches when reference measurement procedures are not available. Collaboration among PT/EQA scheme providers can reduce costs by the sharing of samples and target value assignment to amortize the cost over larger numbers of participants. Globally relevant PT/EQA summary reports would be valuable to advance the practice of laboratory medicine. Residual commutable samples can be used by IVD manufacturers as part of their internal calibration procedures, and could be supplied to laboratories and manufacturers for validation of new or existing measurement procedures.

PT/EQA providers are in a unique position to add substantial value to the practice of laboratory medicine by identifying analytes that are in need of standardization or harmonization, and by stimulating and sustaining global standardization and harmonization initiatives that are needed to support clinical practice guidelines.

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

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: W. Greg Miller, *Clinical Chemistry*, AACC, and CLSI.
Consultant or Advisory Role: Graham R.D. Jones, Royal College of Pathologists of Australasia Quality Assurance Programs Pty Ltd.; W. Greg Miller, College of American Pathologists.
Stock Ownership: None declared.
Honoraria: Gary L. Horowitz, College of American Pathologists.
Research Funding: None declared.
Expert Testimony: None declared.
Other: Graham R.D. Jones, Royal College of Pathologists of Australasia Quality Assurance Programs Pty Ltd.; Gary L. Horowitz, College of American Pathologists.

References

1. Belk WP, Sunderman FW. A survey of the accuracy of chemical analyses in clinical laboratories. Am J Clin Pathol 1947;17:853–61.
2. Wootton ID, King EJ. Normal values for blood constituents; inter-hospital differences. Lancet 1953;1:470–1.
3. CLSI. Using proficiency testing to improve the clinical laboratory; approved guideline. 2nd ed. CLSI document GP27–A2. Wayne (PA): CLSI; 2007.
4. International Organization for Standardization/International Electrotechnical Commission. Conformity assessment: general requirements for proficiency testing. ISO 17043. Geneva: ISO/IEC; 2010.
5. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52:553–4.
6. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. Clin Biochem Rev 2007;28:139–47.
7. CLSI. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline. CLSI document C53-A. Wayne (PA): CLSI; 2010.
8. International Organization for Standardization. In vitro diagnostic medical devices—measurement of quantities in biological samples—metrological traceability of values assigned to calibrators and control materials. ISO 17511. Geneva: ISO; 2003.
9. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. Clin Chim Acta 2003;327:25–37.
10. Miller WG, Myers GL, Ashwood ER, Killeen AA, Wang E, Thienpont LM, Siekmann L. Creatinine measurement: state of the art in accuracy and inter-laboratory harmonization. Arch Pathol Lab Med 2005;129:297–304.
11. Schreiber WE, Endres DB, McDowell GA, Palomaki GE, Elin RJ, Klee GG, Wang E. Comparison of fresh frozen serum to proficiency testing material in College of American Pathologists surveys: α-fetoprotein, carcinoembryonic antigen, human chorionic gonadotropin, and prostate-specific antigen. Arch Pathol Lab Med 2005;129:331–7.
12. Bock JL, Endres DB, Elin RJ, Wang E, Rosenzweig B, Klee GG. Comparison of fresh frozen serum to traditional proficiency testing material in a College of American Pathologists survey for ferritin, folate, and vitamin B12. Arch Pathol Lab Med 2005;129:323–7.
13. Miller WG, Myers GL, Ashwood ER, Killeen AK, Wang E, Ehlers GW, et al. State of the art in trueness and inter-laboratory harmonization for 10 analytes in general clinical chemistry. Arch Pathol Lab Med 2008;132:838–46.
14. CLSI. Preparation and validation of commutable frozen human serum pools as secondary reference materials for cholesterol measurement procedures; approved guideline. CLSI document C37-A. Wayne (PA): CLSI; 1999.
15. Cobbaert C, Weykamp C, Baadenhuijsen H, Kuypers A, Lindemans J, Jansen R. Selection, preparation, and characterization of commutable frozen human serum pools as potential secondary reference materials for lipid and apolipoprotein measurements: study within the framework of the Dutch project "Calibration 2000'." Clin Chem 2002;48:1526–38.
16. National Kidney Disease Education Program. Laboratory professionals commutability study of creatinine reference materials. http://www.nkdep.nih.gov/labprofessionals/commutabilitystudy.htm (Accessed July 2011).
17. Miller WG, Thienpont LM, Van Uytfanghe K, Clark PM, Lindstedt P, Nilsson G, Steffes MW. Toward standardization of insulin immunoassays. Clin Chem 2009;1011–8.
18. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 1: thyroid-stimulating hormone. Clin Chem 2010;56:902–11.
19. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 2: free thyroxine and free triiodothyronine. Clin Chem 2010;56:912–20.
20. Thienpont LM, Van Uytfanghe K, Beastall G, Faix JD, Ieiri T, Miller WG, et al. Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 3: total thyroxine and total triiodothyronine. Clin Chem 2010;56:921–9.
21. Little RR, Rohlfing CL, Tennill AL, Madsen RW, Polonsky KS, Myers GL, et al. Standardization of C-peptide measurements. Clin Chem 2008;54:1023–6.
22. Thienpont LM, Van Uytfanghe K, Marriott J, Stokes P, Siekmann L, Kessler A, et al. Feasibility study of the use of frozen human sera in split-sample comparison of immunoassays with candidate reference measurement procedures for total thyroxine and total triiodothyronine measurements. Clin Chem 2005;51:2303–11.
23. Miller WG. Matrix effects in the measurement and standardization of lipids and lipoproteins. In: Rifai N, Warnick GR, Dominiczak MH, eds. Handbook of lipoprotein testing. 2nd ed. Washington (DC): AACC Press; 2000. p 695–716.
24. Howanitz JH. Review of the influence of polypeptide hormone forms on immunoassay results. Arch Pathol Lab Med 1993;117:369–72.
25. Satterfield MB, Welch MJ. Comparison by LC-MS and MALDI-MS of prostate-specific antigen from

five commercial sources with certified reference material 613. Clin Biochem 2005;38:166–74.
26. Bureau International des Poids et Mesures. Joint Committee for Traceability in Laboratory Medicine. Database of higher-order reference materials, measurement methods/procedures and services. http://www.bipm.org/jctlm/ (Accessed July 2011).
27. Blirup-Jensen S, Johnson AM, Larsen M. Protein value transfer: a practical protocol for the assessment of serum protein values from a reference material to a target material. Clin Chem Lab Med 2008;46:1470–9.
28. Broughton PM, Eldjarn L. Methods of assigning accurate values to reference serum; part 1: the use of reference laboratories and consensus values, with an evaluation of a procedure for transferring values from one reference serum to another. Ann Clin Biochem 1985;22:625–34.
29. Eldjarn L, Broughton PM. Methods of assigning accurate values to reference serum; part 2: the use of definitive methods, reference laboratories, transferred values and consensus values. Ann Clin Biochem 1985;22:635–49.
30. Department of Health and Human Services. Centers for Disease Control and Prevention. Current CLIA regulations. http://wwwn.cdc.gov/clia/regs/toc.aspx (Accessed July 2011).
31. Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. www.bundesaerztekammer.de/downloads/Rili-BAeK-Labor.pdf (Accessed July 2011).
32. Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Strategies to set global analytical quality specifications in laboratory medicine: consensus agreement. Scand J Clin Lab Invest 1999;59:475–585.
33. Ross JW, Fraser MD. Analytical goals developed from the inherent error of medical tests. Clin Chem 1993;39:1481–93.
34. Witte DL. Medically-relevant laboratory performance goals: a listing of the complexities and a call for action. Clin Chem 1993;39:1530–5.
35. Ricos C, Baadenhuijsen H, Libeer CJ, Petersen PH, Stockl D, Thienpont L, Fraser CC. External quality assessment: currently used criteria for evaluating performance in European countries, and criteria for future harmonization. Eur J Clin Chem Clin Bichem 1996;34:159–65.
36. Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76–83.
37. Bais R. What information should manufacturers provide on their procedures? Clin Chem 2006;52:1624–5.
38. Singh RJ, Grebe SK, Yue B, Rockwood AL, Cramer JC, Gombos Z, et al. Precisely wrong? Urinary fractionated metanephrines and peer-based laboratory proficiency testing. Clin Chem 2005;51:472–4.
39. Ross JW, Miller WG, Myers GL, Praestgaard J. The accuracy of laboratory measurements in clinical chemistry: a study of eleven routine analytes in the College of American Pathologists Chemistry Survey with fresh frozen serum, definitive methods and reference methods. Arch Pathol Lab Med 1998;122:587–608.
40. Miller WG, Myers GL, Gantzer ML, Kahn SE, Schönbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
41. Miller WG. Estimating glomerular filtration rate. Clin Chem Lab Med 2009;47:1017–9.
42. Little RR, Rohlfing CL, Sacks DB for the National Glycohemoglobin Standardization Program (NGSP) Steering Committee. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. Clin Chem 2011;57:205–14.
43. Miller WG, Ross JW. The combined target approach: a way out of the proficiency testing dilemma. Arch Pathol Lab Med 1994;118:775–6.
44. Baadenhuijsen H, Weykamp C, Kuypers A, Franck P, Jansen R, Cobbaert C. Commuteerbaarheid van het huidige monstermateriaal in de SKML-rondzendingen van de algemene klinische chemie. Ned Tijdschr Klin Chem Labgeneesk 2008;33:154–7.
45. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232–5.