

The role of proficiency testing in achieving standardization and harmonization between laboratories[☆]

W. Greg Miller^{*}

Virginia Commonwealth University, Richmond, VA, USA

Received 1 September 2008; accepted 1 September 2008

Abstract

Objectives: Proficiency testing is part of a quality system that provides external assessment that a laboratory's performance meets clinical requirements.

Design and methods: A set of samples, intended to simulate clinical samples, is provided to a group of laboratories. The samples are measured for specified measurands and the results reported to the organizing body. Target values are assigned by the organizing body typically as the mean response of a group of laboratories that use similar measurement technology; or, when samples are commutable with native clinical samples, by a reference measurement procedure.

Results: Each participant's results are compared to the target values to determine if a laboratory's results are in conformance to the expectations for clinically acceptable results. When commutable samples are used, results can be compared to reference measurement procedure values to evaluate trueness. Unacceptable results are investigated and corrective action taken.

Conclusion: Proficiency testing allows an individual laboratory to evaluate if its performance is satisfactory for the intended clinical purpose. Comparison of results among laboratories allows evaluation of the effectiveness of calibration procedures in the field.

© 2008 The Canadian Society of Clinical Chemists. Published by Elsevier Inc. All rights reserved.

Keywords: Proficiency testing; Standardization; Harmonization

Introduction and overview of proficiency testing (PT)

Proficiency testing is an external assessment of a laboratory's analytical performance in comparison to its peers or to an accuracy based reference system. PT is also referred to as external quality assessment (EQA) in some countries and is intended to include assessment of pre- and post-examination practices in addition to analytical performance [1]. PT is a component of a laboratory's total quality system that is intended to verify on a periodic, but recurring, basis that laboratory results conform to the expectations for quality set by the organizing body. This report will only address PT for quantitative measurement procedures.

A PT survey is conducted by sending a set of samples from an organizing body to a participating laboratory for measure-

ment by that laboratory in the same manner as is usually used to measure clinical samples. The PT samples are intended to simulate the native clinical samples usually measured by the laboratory. Results for the set of samples are returned to the PT organizer for evaluation of conformance to the results expected for those samples. A report is prepared by the organizer that includes the results reported by a laboratory, the method used for the measurements, the target values expected for each measurand in each sample, the distribution of results from all participants in the survey, and an evaluation if each result reported is acceptable. PT organizers also include summary statistics to describe performance of the group of participants.

Two general procedures are used to assign target values. The most common procedure is to categorize participant methods into "peer groups" that represent similar technology that is likely to achieve the same result and calculate the mean value of the peer group as the target value. Peer grouping is necessary because most sample materials used for PT are modified during preparation in such a manner that the matrix is altered relative to native clinical samples and the PT samples are not commutable with native

[☆] Presented at the First International Symposium on Quality and Accreditation in Laboratory Medicine, Istanbul, Turkey, 16 April 2008.

^{*} 401 N. 13th Street, Room 501, Richmond, VA 23298-0286, USA. Fax: +1 804 828 0353.

E-mail address: gmillier@vcu.edu.

clinical samples (discussed in more detail later). Consequently, different measurement methods have different magnitudes of matrix related bias and do not give the same numeric result for a PT sample. It is common that peer groups are formed as instrument/method groupings from the same manufacturer.

The second procedure to assign a target value is by measurement of a PT sample using a reference measurement procedure. This approach can be used when the PT material is commutable with native clinical samples. A commutable PT sample is one that has an equivalent mathematical relationship as that observed for native clinical samples between all the different measurement procedures represented in the survey. It is uncommon that PT samples are commutable with native clinical samples, consequently target value assignment with a reference measurement procedure is only used in special cases.

The evaluation criteria for agreement of a participant result with the target value are established by the organizer based on either a statistical distribution (e.g. ± 3 SD) of participant results around a peer group target value, or by an allowable difference that may be a percent or a fixed concentration difference (e.g. 10% or 6 mmol/L whichever is larger). Ideally, an allowable difference will be based on clinically relevant criteria, but in many cases the acceptance criteria will need to also consider the influence of matrix bias and between laboratory variability. In some countries, the criteria are set by regulatory or accreditation requirements.

Evaluation of PT results

Practical guidance for clinical laboratories regarding use and evaluation of PT results is provided by the Clinical and Laboratory Standards Institute document, *Using Proficiency Testing to Improve the Clinical Laboratory* [1]. The CLSI document includes guidance on selection of a PT program, PT sample handling procedures, the evaluation and reporting process, investigation of and response to unsatisfactory scores, and monitoring PT performance over time. The guidance also covers assessment of pre- and post-examination phases of laboratory testing, and using PT as an educational tool.

Considerations in selecting a PT program include:

- Traceable reference values (when samples are commutable and reference values are used);
- Samples that behave like patient samples (commutable) when possible;
- Method compatibility with available peer groups;
- Size of available peer groups;
- Frequency of challenges;
- Timeliness and usefulness of reports;
- Educational content;
- Customer service.

Appropriate sample handling considerations include:

- PT samples often require some additional preparatory manipulations that are not used for the usual clinical samples (this step may be a source of error);
- Otherwise, PT samples must be tested in the same manner as patient samples (e.g. singlet if patients are measured in singlet);

- PT testing should be rotated among all the personnel who usually perform patient testing;
- PT samples can be retained for future follow up testing with storage consistent with the measurand stability (typically frozen at -70°C or lower for fluids).

Each unacceptable PT result needs to be investigated and the findings and corrective action documented. It is good practice that a laboratory follow up on results that may not have been scored as unsatisfactory by the provider, but that may represent a statistically unlikely result and a potential error condition for the measurement system. For example, it may be useful to do a formal investigation for results that are >2.5 SD from the target value. Steps in the investigation include:

- Gather data related to the testing event to include review of records of calibration, quality control, reagent use, maintenance procedures, etc.;
- Identify the root cause of the error;
- Take corrective action;
- Monitor the success of the corrective action;
- Document the investigation and corrective action;
- Report the results of the investigation to the oversight body if requested or if required by regulations.

The CLSI guideline suggests that problems identified when investigating unacceptable PT results can be classified as follows. This classification scheme allows a laboratory to develop an appropriate corrective action plan.

1. Clerical errors:

- Incorrectly transcribed PT result from the instrument read-out to the report form;
- The PT sample was mislabeled in the laboratory;
- An incorrect instrument or method was reported on the results submission form;
- Incorrect units were reported;
- A decimal point was misplaced.

2. Methodologic problems:

- Inadequate standard operating procedure (SOP);
- 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;
- Carry-over from a previous sample;
- Inadequate QC procedures were 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 malfunction;
 - 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 preparation or storage of reagents, calibrators, PT materials (e.g. after opening);
 - 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:
- Differences between PT samples and patient samples, e.g. matrix, additives, stabilizers;
 - Sample deteriorated in transit;
 - Sample had weak or borderline reaction;
 - Sample contained interfering factors (which may be method-specific);
 - Sample was not homogeneous among vials.

If no explanation can be identified after investigation, the laboratory concludes that a random error occurred. It is important to recognize that a single random erroneous result is not conclusive and that corrective action may not be indicated. Attempting to take corrective action when a root cause cannot be identified may introduce an error when the original unacceptable result was a non-persistent random event.

It is common practice to repeat the measurement for a stored aliquot of the PT sample that had an unacceptable result (assuming the measurand was stable on storage) as well as the other PT samples from the same set to confirm that corrective action was successful or to confirm that the original unacceptable result was a random event. If the repeated result is still unacceptable, the laboratory concludes that a systematic error is present and conducts further investigation. In reviewing the erroneous result, the other results from the set of PT samples should also be considered for a pattern of random or systematic errors. Multiple results with relatively large differences scattered on both sides of the target value suggests inadequate method precision. Multiple results with relatively large differences in the same direction from the target value suggests a bias problem.

It is recommended to monitor trends in results for PT samples from different PT events over time. It is common to have multiple PT events during a year. A graphical representation of results (as number of SD or percent different than the target value) over a one or more years time can allow a laboratory to identify that a method has a systematic bias even though results may not have individually been scored as unacceptable. In such a case, the laboratory can initiate corrective action before an erroneous condition reaches a critical situation.

Limitations in interpreting PT results

The principle limitation in interpreting PT results is the influence of non-commutable samples on the results. As stated earlier, a

commutable PT sample is one that has an equivalent mathematical relationship as that observed for native clinical samples between all the different measurement procedures represented in the survey. Commutability and its influence on PT and other reference materials has been reviewed [2–4] and will not be discussed in detail here. Numerous investigations have reported that the surrogate patient sample materials typically used as PT samples are not commutable with native clinical samples [2–8]. Non-commutability is caused by manipulation and alteration of the sample matrix during preparation of the PT material even if the PT material is based on or derived from materials that were collected from human subjects.

Consequently, it is not possible to determine if the difference observed between two or more measurement procedures (including a reference measurement procedure) is caused by a calibration difference or a bias caused by an altered matrix. Because of the non-commutability limitation, PT results from such materials must be compared to the mean of results from participants that use measurement procedures that are expected to have the same or very similar matrix related bias. Instrument and method combinations are categorized into “peer groups” that use the same or similar measurement conditions so the results are likely to be comparable to each other. When non-commutable samples are used, it is not appropriate to compare an individual laboratory’s result to the mean value for a different method peer group, to an all method mean or to the result from a reference measurement procedure. It is also not appropriate to compare the mean result for one method type to the mean result of a different method type.

Peer group evaluation provides very useful information to a participant and can provide valuable information to improve quality in a laboratory. Participant bias vs. the peer group mean value verifies that a laboratory is using a measurement procedure in conformance with the manufacturer’s specifications and in conformance to other laboratories using the same technology. In this situation, trueness of the measurement procedure is provided by the manufacturer of the measurement procedure. Consequently, it is important to use the measurement procedure according to the manufacturer’s instructions.

Peer group evaluation also allows the imprecision within a peer group to be evaluated and participants can understand the effectiveness of a manufacturer’s quality system to deliver uniform results among the group of users of that technology. PT programs may also include the same sample in multiple survey events from which evaluation of the longer-term precision (reproducibility) for a laboratory can be determined.

Accuracy (trueness) based PT programs

There are a number of PT programs in several countries that use commutable samples that are prepared from freshly collected and minimally processed human samples. These programs are frequently limited to one or a small number of measurands that can be obtained from freshly collected human blood or urine donors. One of the oldest such programs is the Glycohemoglobin Survey from the College of American Pathologists in the USA. This survey uses fresh whole blood

collected from healthy and diabetic donors that is pooled, aliquoted, shipped on cold packs, and measured by participants within a few days of blood collection. The target values have been established by reference measurement procedures and the results have been used by the National Glycohemoglobin Standardization Program (USA) as surveillance of the success of manufacturers to achieve standardization of hemoglobin A1c measurement procedures.

PT programs that use commutable samples provide information for participants and also provide information for manufacturers and professional standardization organizations on the traceability of methods to reference systems, on the effectiveness of a manufacturer's transfer of calibration traceability to routine measurement procedures in the field, and on the harmonization or standardization among routine measurement procedures from different manufacturers [5–8].

References

- [1] Clinical and Laboratory Standards Institute (CLSI). *Using Proficiency Testing to Improve the Clinical Laboratory; Approved Guideline—Second Edition*. CLSI document GP27-A2 (ISBN 1-56238-632-8). Clinical and Laboratory Standards Institute, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2007.
- [2] Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. *Clin Chim Acta* 2003;327:25–37.
- [3] Miller WG, Myers GL, Rej R. Why commutability matters. *Clin Chem* 2006;52:553–4 (editorial).
- [4] Vesper HW, Miller WG, Myers GL. Reference materials and commutability. *Clin Biochem Rev* 2007;28:139–47.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] Miller WG, Myers GL, Ashwood ER, Killeen AK, Wang E, Ehlers GW, Hassemer D, Lo SF, Seccombe D, Siekmann L, Thienpont LM, Toth A. State of the Art in Trueness and Inter-Laboratory Harmonization for 10 Analytes in General Clinical Chemistry. *Arch Pathol Lab Med* 2008; in press.