Discussion

Important information regarding grading of specimen C-02

Specimen C-02 was a carefully prepared fresh frozen human serum sample that was sent to all participants in the Survey. Your laboratory's performance with this sample was graded using two separate grading schemes. The first is the usual grading system that compares your result to that obtained by other laboratories using the same method and/or instrument (i.e., peer group grading). The second grading system is used for educational purposes only and will not be reported to regulatory agencies. It is provided so that you can see how your laboratory's results compare to those of all other participants and to reference measurement procedures. Sample C-02 is designed to be free of matrix effects, and therefore results from different laboratories using different instruments should be directly comparable to each other and to the reference method results.

If your laboratory passes the usual grading system, but scores a failing grade with the educational grading system, only the passing score will be used for accreditation purposes.

Evaluation of Survey results for regulatory grading

The main objective of proficiency testing (PT) is to evaluate laboratory performance for clinically acceptable results compared to peer laboratories. Over 7000 laboratories subscribe to the CAP Chemistry Surveys Program. Results from these laboratories are used to create a powerful database for interlaboratory comparisons. The data summary collects individual results into method specific peer groups so that participants can evaluate the quality of their results. Peer group evaluation allows a laboratory to confirm it is using a measurement technology correctly and is producing results in agreement with other users of the same or similar methods. The peer group standard deviations provide information on the relative imprecision of different methods, and on the uniformity of the method manufacturer's calibration process within the peer group.

In pursuit of the College's ongoing mission to develop new ways to provide value and education to the participating laboratories, this Survey included a specimen, C-02, which was specially prepared to be very similar to a freshly collected serum as is usually assayed in the clinical chemistry laboratory.

Individual laboratory results for all Survey specimens (including C-02) have been evaluated in the usual way, using peer group target values and CLIA limits, as described at the beginning of the Participant Summary Report (PSR). This normal evaluation appears for each analyte in the first section of the laboratory evaluation report. The CAP Laboratory Accreditation Program will only review results from the peer group evaluation to monitor laboratory performance, and only this evaluation will be reported to the Centers for Medicare and Medicaid Services (CMS) in compliance with CLIA regulations, if a laboratory has supplied a CLIA Identification Number to the College.

Discussion

Glossary of key terms

Accuracy: the agreement between a method's result and an accepted value for the analyte. Accuracy, as used by the International Organization for Standardization (ISO), is the agreement between a single measurement and the best available value for the analyte content. Accuracy includes bias and imprecision components. A related term, trueness, is defined by ISO as the agreement between the mean of replicate measurements made by a routine method and a reference measurement procedure. In this text, the more common term accuracy is used for agreement between an individual laboratory's result, or a method peer group mean result, and a reference measurement procedure value.

Commutability: an attribute of a PT material in which the result for a PT specimen will have a numeric value which is nearly identical to that expected for a clinical specimen measured by the same method. A PT material having this attribute is said to be commutable for the specified method or methods.

Harmonization: the agreement between numeric results for two different methods. Methods that produce the same numeric results for clinical specimens are said to be harmonized.

Commutability issues in Proficiency Testing

Matrix interferences are generally present in the materials used by all PT providers at the present time. Matrix interferences are caused by alteration of the specimen matrix during the processing steps in material manufacturing. Briefly, plasma collected at donor centers is converted to serum in large batches, which are typically dialyzed to remove impurities; and then various analyte concentrates, including non-human components, are added back to achieve the range of values needed to challenge methods at different concentrations. As a result, a PT specimen may not be commutable among all the methods used for routine analysis. Commutable means the result for a PT specimen will have a numeric value that is nearly identical to that expected for a clinical specimen measured by the same method. Recent literature reports, including studies by the CAP, have documented that processed PT materials are frequently non-commutable and the occurrence of non-commutability has been unpredictable for any particular material/method combination.[1,2]

Non-commutability prevents direct comparison of results between methods. When PT results using non-commutable materials are evaluated, the observed difference between a result and a reference measurement procedure (RMP) target value, or between the peer group means for two different methods, has contributions from calibration bias (trueness or accuracy), random bias, and matrix bias. The calibration bias and random bias are test procedure attributes

Discussion

that reflect performance for patient specimens. Matrix bias is the component of the observed difference due to non-commutability between a method/PT material combination. The presence and magnitude of a matrix bias is typically unknown but adds to the sum of calibration and random bias. Consequently, the total observed difference for a PT material between an individual method result and a RMP target value, or between the peer group means for two different methods, can produce an incorrect inference of test procedure performance for native clinical specimens.[1-4]

Consider the following example from reference 1 to illustrate the differences between commutable and non-commutable PT materials. The difference between an individual laboratory result and a RMP result for a PT specimen is 20 mg/dL. The 20 mg/dL difference includes contributions from calibration bias, random bias and matrix bias. If the laboratory's random bias is known to have a 2 SD value of 5 mg/dL, it would be useful if one could conclude the method has a calibration bias of 15 mg/dL which should be corrected. However, because of a possible matrix bias with the PT material the amount of the 15 mg/dL due to calibration bias cannot be discriminated from the amount due to matrix bias. Thus, the laboratory cannot use the PT result to determine the calibration status of the method when clinical specimens are being measured.

Peer group grading is the method of choice used to evaluate PT results when matrix interferences may be present in the Survey materials. Peer group grading compares an individual result to the mean of all results performed with the same method. This approach assumes any method/material matrix bias is the same for all members of the peer group. Peer group evaluation provides very useful information on the relative accuracy of an individual laboratory by confirming that a measurement technology is used correctly to achieve agreement with other users of the same method.

Currently, most PT specimen materials are not designed to be commutable. The volumes needed and costs associated with manufacturing of the materials preclude preparation of large quantities of fresh serum-based materials. However, there has been increasing interest in use of pooled native clinical specimens for commutable materials in PT programs. The College has used native pooled whole blood in its Glycohemoglobin (GH2) Survey for several years. Some specialized European PT programs have reported successful use of native serum specimens.[4-7] Specimen C-02 was prepared as a very high quality pooled serum specimen to evaluate method performance in this large CAP Survey program.

Discussion

Commutability attributes of Specimen C-02

Specimen C-02 was prepared by a modification of the NCCLS C37-A Guideline.[8] This stringent protocol required donor blood collection and clotting under carefully controlled conditions to yield a serum product that had very little contact time with red blood cells and was frozen within about eight hours of blood collection. The frozen serum units were thawed, pooled, aliquotted and refrozen without any manipulation or additives so that the final pooled serum material is very similar to a normal patient's serum.

Consequently, the results for the C-02 specimen are expected to be free of matrix interferences and commutable among all routine methods and reference measurement procedures. One important goal of clinical laboratory testing is to produce harmonized results that are the same irrespective of the method used. In addition, when a RMP is available, results should agree with the RMP, which ensures both accurate and harmonized results among laboratories. The commutability attribute has allowed specimen C-02 to be evaluated for both accuracy and harmonization of results.

Educational evaluation of Survey results for specimen C-02

The "dual grading" comparison uses a reference measurement procedure (RMP) assigned target value. Because specimen C-02 is commutable, a method's result can be compared to a RMP target value to evaluate the accuracy that would be applicable for a typical patient's serum specimen. The evaluation limits used for "dual grading" of specimen C-02 are the same as used for the usual peer group evaluation as listed on page 1 of the Participant Summary Report. The only difference is the RMP target value was used instead of the peer group mean value. The RMP used for each analyte is shown in Table 1 on the following page. This second evaluation is provided for **educational purposes only** and is not reported to CMS, nor will it be used by the CAP Laboratory Accreditation Program.

Discussion

Table 1. Reference measurement procedures and target values for Specimen C-02.

Analyte	RMP Method	No. RMP Labs	RMP Target Value	RMP Target Value SEM[a]
Albumin	Bromcresol green	2	4.11 g/dL	0.011 g/dL
Bilirubin, total	Jendrassik-Grof	5	0.36 mg/dL	0.015 mg/dL
Calcium	Atomic absorption	3	9.20 mg/dL	0.032 mg/dL
Chloride	Mass spectrometry	1[b]	103.6 mmol/L	0.00 mmol/L
Cortisol	Mass spectrometry	2	13.55 µg/mL	0.067 µg/mL
Creatinine	Mass spectrometry	2	0.90 mg/dL	0.003 mg/dL
Glucose	Mass spectrometry	1[b]	98.6 mg/dL	0.17 mg/dL
Iron	Ferrozine	3	65.4 µg/dL	0.90 µg/dL
Magnesium	Atomic absorption	1	1.90 mg/dL	0.000 mg/dL
Phosphorus	Ammonium molybdate	3	3.25 mg/dL	0.016 mg/dL
Potassium	Flame photometry	3	4.38 mmol/L	0.008 mmol/L
Protein, total	Biuret	4	7.13 mg/dL	0.014 mg/dL
Sodium	Flame photometry	3	140.7 mmol/L	0.22 mmol/L
Thyroxine (T4), total	Mass spectrometry	2	6.43 µg/dL	0.038 µg/dL
Urea (BUN)	Mass spectrometry	1[b]	12.2 mg/dL	0.012 mg/dL
Uric acid	Mass spectrometry	2	5.34 mg/dL	0.017 mg/dL

[a] SEM is standard error of the mean.

[b] In the case of chloride, glucose, and urea only one reference laboratory used mass spectrometry. Mass spectrometry is considered the RMP of higher order, but for many analytes other RMPs are recognized and accepted. Chloride was also assayed by four laboratories using an amperometric RMP with a mean of 104.1 and SEM of 0.51 mmol/L; glucose was assayed by three laboratories using a hexokinase RMP with a mean of 98.3 and SEM 0.32 mg/dL; and urea (BUN) was assayed by two laboratories using a urease RMP with a mean 12.2 and SEM 0.08 mg/dL.

Laboratory directors need to use caution when interpreting the educational evaluation vs. a RMP target value. A miscalibrated method can be identified as a bias by an individual laboratory vs. the RMP. However, the miscalibration could be due to the individual laboratory not applying the method correctly, or due to the method manufacturer not having a robust calibration process. The peer group evaluation for the Survey specimens can be used to determine that a laboratory is applying the method correctly. For many analytic systems, the method manufacturer provides the instrument, reagents and calibrators. In these cases, an individual laboratory cannot easily correct a method calibration bias; rather the method manufacturer needs to evaluate the method calibration status.

Discussion

For example, the following analytes had one or more method peer groups which had higher rates of disagreement, relative to other peer groups, between individual participant results and the RMP target value: albumin, cortisol, magnesium, phosphorous, thyroxine, and sodium. Brief comments follow for each analyte based on review of the Participant Summary Report peer group mean values and standard deviations (SD) vs. the RMP target values.

Albumin peer group mean values show four method groups (Abbott Aeroset, Roche Cobas Fara/Mira, Olympus, and Vitros) that have higher mean biases than the others. These higher mean biases contributed to the number of individual participants who were flagged as outside the educational evaluation limits (77% for Aeroset, 30-44% for Fara/Mira depending on reagent, 3% for the Olympus 400-640/2700/5400 group, and 4-8% for Vitros depending on model). In this situation, the individual participant is dependent on the manufacturer to review the method calibration process. An important consideration for those methods that are "open", i.e. reagents and calibrators can be obtained from several sources, is that the observed peer group mean and SD could be skewed by one or more reagent systems that are not under the control of the instrument manufacturer. In such a case, individual laboratories should evaluate the suitability of their choice of vendor for reagent and calibrator.

Cortisol is an analyte which had overall good performance with the peer group mean biases, generally within a range of ± 2 μg/dL of the RMP target value. However, three peer groups (Bayer ACS 180 and Advia Centaur; Tosoh AIA-PACK) with mean biases near 2 μg/dL had 23%, 15%, and 9%, respectively, of individual participant biases outside the evaluation limit of approximately 4 μg/dL. The individual laboratories with higher values had an excessive total error from the combination of the 2 μg/dL bias and imprecision of results reflected in the peer group standard deviations (1.8, 1.5, and 1.4 μg/dL, respectively). On the other hand, the Vitros ECI had a peer group mean bias of -2 μg/dL with 0% of individual participants exceeding the evaluation limit due to the small imprecision (0.5 μg/dL) for this group, which kept the total error within the evaluation criterion. Imprecision can come from several sources including calibration frequency, instrument maintenance, reagent storage and handling, lot-to-lot consistency of reagents and calibrators, etc. Laboratories should review these parameters with assay system manufacturers.

Magnesium had small peer group mean biases vs. the RMP ranging from -0.1 to 0.2 mg/dL for all but one method. The Abbott Aeroset had a peer group mean bias of 0.4 mg/dL, which caused 21% of individual participants to exceed the 0.5 mg/dL evaluation criterion. This peer group is the only one to use an arsenazol based reagent. Whether the bias is due to non-specificity of the reagent or a miscalibration is not possible to determine from this data. In any event, individual laboratories are dependent on the reagent and calibrator manufacturer to correct the method bias.

Discussion

Phosphorous had peer group mean biases vs. the RMP from −0.19 to 0.23 mg/dL, which represents overall good agreement among methods. However, the evaluation criterion of 0.4 mg/dL is sufficiently stringent that five peer groups (three Beckman Synchron and two Roche Cobas systems using various reagents) had total error large enough that 6-12% of individual participants did not meet the criterion. The total error is the combination of bias and imprecision that affects the distribution of results for a peer group. As has been mentioned, the observed peer group performance for methods such as these that are "open" may be skewed by one or more reagent systems, or use of reagents and calibrators from different manufacturers, that are not under the control of the instrument manufacturer. The individual laboratory should review the suitability of reagents and calibrators for the instrument.

Thyroxine (T4) had five peer groups (Bayer ACS 180 and Advia Centaur; Beckman Synchron Syst; Cedia; Microgenics DRI) with mean biases vs. the RMP between 0.67 and 0.96 μg/dL, which had 8-14% of individual participants' results exceed the educational evaluation criterion of 1.3 μg/dL. Standardization for thyroxine measurements is a recognized need and programs are in development to improve calibration harmonization for this analyte.

Several undiluted sodium methods had peer group mean biases large enough to cause a relatively high percent of individual participant biases vs. the RMP target value to exceed the evaluation criterion. All of the diluted sodium methods had peer group mean biases between −0.6 and 1.4 mmol/L, whereas the undiluted methods biases ranged from 0.7 to 6.0 mmol/L. Three peer groups (Dade Behring Dimension/AR/MULT; Nova CRT series; Roche Cobas FARA/MIRA) had mean biases of about 5-6 mmol/L with 48%, 88%, and 76%, respectively, of individual participants' biases exceeding the evaluation criterion of 4 mmol/L. These large calibration biases account for the entire range of the evaluation limit. The Vitros peer groups had smaller mean biases ranging from 2-3 mmol/L. However, this bias was large enough to cause 6-17%, depending on peer group, of individual participants to exceed 4 mmol/L from the RMP value. The Roche Cobas Integra had a small 1 mmol/L mean bias but 12% of individual participants' biases vs. the RMP exceeded the evaluation criterion due to a greater dispersion of values for this method as reflected in the standard deviation. Because ISE methods should only use reagents and calibrators from the instrument manufacturer, these calibration biases should be addressed by the method manufacturers.

Evaluation of harmonization for method peer groups

The state of the art in harmonization of results between method groups can be evaluated from the peer group mean results for Specimen C-02. The Participant Summary Report provides detailed information on the performance of each peer group. Table 2, on the following page, summarizes the status of harmonization and accuracy for the 16 analytes that have RMP target values. The choice of 10% is arbitrary as a criterion for agreement with the RMP value, but provides a basis for comparison. There is generally good agreement in results among the various method groups used by laboratories for these 16 analytes with exception of bilirubin and creatinine, plus the specific peer groups discussed in the preceding section (albumin, cortisol, magnesium, phosphorous, thyroxine, and sodium).

Discussion

Bilirubin had 20 of 36 peer groups with mean values greater than 10% from the RMP target value. However, the target value of 0.36 mg/dL is a low value consistent with a non-supplemented human sera pool. The peer group standard deviations ranged from 0.04 to 0.19 mg/dL, which coupled with the bias ranges, suggests the total errors for these routine methods at this low value may not be clinically significant.

Creatinine had 17 of 42 peer group mean biases greater than 10% of the RMP value of 0.9 mg/dL. These peer groups had biases vs. the RMP typically between 0.1-0.2 mg/dL. This magnitude is clinically significant particularly when creatinine is used to calculate a value for glomerular filtration rate as recommended by the NIH National Kidney Disease Education Program (NKDEP; http://www.nkdep.nih.gov/). Methods using the Jaffe alkaline picrate reagent are known to have non-specific reactions with serum proteins and other metabolites.[9] The degree of interference varies depending on how the method is implemented (e.g., end point vs. kinetic). Thus, there are both reagent non-specificity and calibration issues to be addressed for creatinine methods. The NKDEP has established a laboratory standardization working group to assist manufacturers to address improvement in creatinine method performance.

Table 2. Harmonization and accuracy for 16 analytes with RMP target values.

Analyte	RMP value	Peer group mean bias minimum difference	Peer group mean bias maximum difference	Peer groups with mean bias >10% of RMP value
Albumin	4.11 g/dL	-0.36	0.71	2/28
Bilirubin, total	0.36 mg/dL	-0.04	0.23	20/36
Calcium	9.20 mg/dL	-0.38	0.05	0/27
Chloride	103.6 mmol/L	-1.18	3.23	0/27
Cortisol	13.55 μg/mL	-2.07	2.47	4/10
Creatinine	0.90 mg/dL	-0.05	0.21	17/42
Glucose	98.6 mg/dL	-1.90	4.71	0/29
Iron	65.4 μg/dL	-7.31	5.19	1/23
Magnesium	1.90 mg/dL	-0.06	0.40	1/25
Phosphorus	3.25 mg/dL	-0.19	0.23	0/26
Potassium	4.38 mmol/L	-0.09	0.13	0/28
Protein, total	7.13 mg/dL	-0.30	0.19	0/22
Sodium	140.7 mmol/L	-0.62	6.04	0/29
Thyroxine (T4), total	6.43 μg/dL	-0.52	0.96	5/18
Urea (BUN)	12.2 mg/dL	-1.59	1.72	4/27
Uric acid	5.34 mg/dL	-0.15	0.41	0/21

Discussion

References:

1. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. *Clin Chim Acta*. 2003;327:25-37.
2. Ross JW, Miller WG, Myers GL, et al. The accuracy of laboratory measurements in clinical chemistry. A study of 11 routine chemistry 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.
3. Libeer JC, Baadenhuijsen H, Fraser CG, et al. Characterization and classification of external quality assessment schemes (EQA) according to objectives such as evaluation of method and participant bias and standard deviation. *Eur J Clin Chem Clin Biochem*. 1996;34:665-678.
4. Thienpont LM, Stockl D, Friedecky B, et al. Trueness verification in European external quality assessment schemes: time to care about the quality of the samples. *Scand J Clin Lab Invest*. 2003;63:195-202.
5. Cobbaert C, Weykamp C, Baadenhuijsen H, et al. 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;48:1526-1538.
6. Stockl D, Libeer JC, Reinauer H, et al. Accuracy-based assessment of proficiency testing results with serum from single donation: possibilities and limitations. *Clin Chem*. 1996;42:469-470.
7. Linko S, Himberg JJ, Thienpont L, et al. Assessment of the state-of-the-art trueness and precision of serum total-calcium and glucose measurements in Finnish laboratories - the QSL-Finland study. *Scand J Clin Lab Invest*. 1998;58:229-240.
8. Preparation and validation of commutable frozen human serum pools as secondary reference materials for cholesterol measurement procedures; approved guideline. NCCLS document C37-A (ISBN 1-56238-392-2). Wayne, PA: NCCLS, 1999.
9. Newman DJ, Price CP. Renal function and metabolites, creatinine and creatine. In: Burtis CA, Ashwood ER, eds. *Tietz Textbook of Clinical Chemistry.* 3rd ed. Philadelpia, PA: Saunders; 1999:1241-1245.