Eline A. E. van der Hagen, Cas Weykamp, Sverre Sandberg, Anne V. Stavelin, Finlay MacKenzie and W. Greg Miller*

Feasibility for aggregation of commutable external quality assessment results to evaluate metrological traceability and agreement among results

https://doi.org/10.1515/cclm-2020-0736
Received May 15, 2020; accepted July 22, 2020; published online August 6, 2020

Abstract

Objectives: External quality assessment (EQA) with commutable samples is used for assessing agreement of results for patients' samples. We investigated the feasibility to aggregate results from four different EQA schemes to determine the bias between different measurement procedures and a reference target value.

Methods: We aggregated EQA results for creatinine from programs that used commutable EQA material by calculating the relative difference between individual participant results and the reference target value for each sample. The means and standard errors of the means were calculated for the relative differences. Results were partitioned by methods, manufacturers and instrument platforms to evaluate the biases for the measurement procedures.

Results: Data aggregated for enzymatic methods had biases that varied from −8.2 to 3.8% among seven instrument platforms for creatinine at normal concentrations (61–85 μmol/L). EQA schemes differed in the evidence provided about the commutability of their samples, and in the amount of detail collected from participants regarding the measurement procedures which limited the ability to sub-divide aggregated data by instrument platforms and models.

Conclusions: EQA data could be aggregated from four different programs using different commutable samples to determine bias among different measurement procedures. Criteria for commutability for EQA samples as well as standardization of reporting the measurement methods, reagents, instrument platforms and models used by participants are needed to improve the ability to aggregate the results for optimal assessment of performance of measurement procedures. Aggregating data from a larger number of EQA schemes is feasible to assess trueness on a global scale.

Keywords: commutability; external quality assessment; harmonization; metrological traceability; standardization.

Introduction

An important goal in laboratory medicine is to have equivalent results for a measurand irrespective of the measurement procedures used in different locations. Equivalent results become increasingly important because more and more clinical guidelines refer to "cut off limits" or "action limits" which are independent of the analytical platform used. The International Organization for Standardization (ISO) standard 17511 describes various calibration hierarchies used by *in vitro* diagnostics (IVD) manufacturers and clinical laboratories to achieve metrological traceability [1] and thus equivalent results for clinical samples. External quality assessment (EQA) has an essential role in surveillance of the continuing success of metrological traceability and harmonization of results among clinical laboratory measurement procedures [2–4].

EQA data from commutable samples is useful to assess the metrological traceability of results from an individual laboratory's measurement procedure to higher order references [2]. In addition to evaluating an individual laboratory's performance, the mean or median of all participant results for a particular measurement procedure (referred to as a peer-group value) has been recommended for

***Corresponding author: W. Greg Miller**, Department of Pathology, Virginia Commonwealth University, PO Box 980286, Richmond, VA, 23298-0286, USA, Phone: +1 804 828 0375; Fax: +1 804 828 0353, E-mail: greg.miller@vcuhealth.org
Eline A. E. van der Hagen and Cas Weykamp, Dutch Foundation for Quality Assessment in Medical Laboratories (SKML), Nijmegen, The Netherlands; Department of Clinical Chemistry, Queen Beatrix Hospital, Winterswijk, The Netherlands
Sverre Sandberg and Anne V. Stavelin, The Norwegian Organisation for Quality Improvement of Laboratory Examinations (Noklus), Haraldsplass Deaconess Hospital, Noklus, Bergen, Norway
Finlay MacKenzie, Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

assessing the metrological traceability of that measurement procedure [2, 3]. The mean or median from aggregated EQA results for one measurement procedure can be compared with that for other measurement procedures to assess the metrological traceability of all measurement procedures included in an EQA scheme. In addition, EQA results from commutable samples can assess the harmonization of results among different measurement procedures by comparing the peer-group median or mean values when there is no value assigned by higher order references. Recent reports have shown the usefulness of EQA for examining performance capability of measurement procedures within a country and among geographically distributed regions. For example, Weykamp et al. shared the same commutable samples in EQA schemes for five countries and showed both agreement and differences in results among manufacturers and countries for 17 general chemistry analytes [4] and for HbA_{1c} in 17 countries [5].

A limitation of using only results from one EQA provider is that there could be a limited number of participants using a measurement procedure that would make the uncertainty of the mean or median value unreliable for use to evaluate its metrological traceability. In addition, different measurement procedures could perform differently in different countries and different regions due, for example, to different calibrations and reagents. Consequently, an approach to aggregate results from different EQA providers can provide larger numbers of results from a given measurement procedure and thus provide a mean or median with low enough uncertainty to evaluate metrological traceability. In addition, useful information on measurement procedure performance over a wide geographic area can be obtained. However, distributing commutable EQA samples, particularly when liquid, to a large number of geographically distributed participants is a challenge due to stability limitations and the volume of material needed.

We report here our experience from a pilot program that aggregated results from four EQA schemes in four countries where samples expected to be commutable and a target value traceable to a reference measurement procedure (RMP) were used. The goals for the pilot were (a) to examine the feasibility to aggregate data from different schemes, (b) to demonstrate how the aggregated data can be used to assess biases among IVD devices used in clinical laboratories, and (c) to determine how to present the data in a form that is useful to the IVD industry and medical laboratories for assessing the effectiveness of harmonization/standardization of measurement procedures. This project is a collaboration between the International Consortium for Harmonization of Clinical Laboratory Results and the European Organization for External Quality Assurance Providers in Laboratory Medicine. The long term goal for the project is to provide global information regarding the status of harmonization/standardization of measurement procedures used in medical laboratories to support maintaining current calibration hierarchies, to improve those calibration hierarchies, or to develop new calibration hierarchies as needed.

Materials and methods

EQA providers

Four EQA providers, the College of American Pathologists (CAP), the Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), the Dutch Foundation for Quality Assessment in Medical Laboratories (SKML), and the United Kingdom National External Quality Assessment Scheme (UK NEQAS), were asked to provide an electronic file of individual participants' results from one distribution that used a serum sample with a normal concentration of creatinine that was expected to be commutable with human samples. Note that creatinine was chosen as an example measurand to examine approaches for aggregation of results. Noklus presents the Norwegian results from a Labquality (Helsinki, Finland) survey. All results were measured in the time interval between October 2018 and January 2019. The file format was at the convenience of the EQA provider and included whatever information was the usual practice regarding the method, reagents, and measurement procedure used by participants. The serum samples were not to include known interfering substances that might have been present to challenge the specificity of some measurement procedures in that distribution. Table 1 describes each EQA scheme. Target values for creatinine were assigned by each EQA provider as indicated in Table 1.

The assigned target value for Noklus was established by transferring the value from the Nordic Society of Clinical Chemistry Reference Serum X (RSX). The certified value of RSX (70.83 μmol/L, expanded uncertainty U=1.13) was established using isotope dilution-gas chromatography-mass spectrometry (ID-GC/MS) as described by Stöckl et al. [6] and Thienpont et al. [7]. The transferred target value of the EQA sample used in the scheme was established as follows: five Nordic laboratories, each using a different measurement procedure, analyzed the EQA sample and the RSX in triplicates. The transferred value (T) for the EQA sample from each laboratory was then calculated as:

$$T = (\text{mean EQA sample from 1 lab}) \times (\text{Certified value for RSX})/ \times (\text{mean of RSX from the 1 lab})$$

The mean of the transferred values from the five laboratories was used as the assigned target value (85.00 μmol/L), and the standard uncertainty was calculated as standard error of this mean (0.5%). The combined expanded uncertainty including the transfer step and the RSX certified value, was U=1.88% (k=2).

Data transformation and aggregation of data

Transformation and merging of data was programmed and executed using R studio. When aggregating results of the four EQA providers,

Table 1: Characteristics of EQA schemes.

EQA provider	CAP	Noklus[a]	SKML	UK NEQAS
Measurement dates	Nov–Dec 2018	Nov 2018	Oct 2018	Jan 2019
Number of participants	336	75	198	402
Sample characteristics	Frozen pooled serum[b]	Frozen pooled serum[c]	Frozen pooled serum[d]	Frozen pooled serum[e]
Commutability assessment	Previous batch in 2006[f]	Not formally assessed	Previous batch in 2005[g]	Not formally assessed
RMP used	IDMS[h]	IDMS transferred value (see text)	IDMS[h]	IDMS[i]
Creatinine value, µmol/L	61.01	85.00	67.87	70.98
Expanded uncertainty	1.1% (k=2.6)[j]	1.88% (k=2)	1.0% (k=2.6)[j]	0.88% (k=2)

[a]Samples were prepared by the Danish Institute for External Quality Assurance for Laboratories and distributed by Labquality, Finland. [b]Prepared according to Clinical and Laboratory Standards C37 protocol. Samples were stored frozen at −70 °C, distributed on cold packs and thawed in transit. [c]Blood was collected into dry blood bags at Herlev Hospital (Denmark) from seven patients with Hemochromatosis and allowed to clot at 4 °C. Serum was separated on the following day, then frozen at −80 °C in donor bags (approximately 200 mL serum). Frozen serum was stored at −80 °C prior to thawing, pooling, filtration and aliquoting. The aliquots were again stored at −80 °C until shipment on dry ice to Labquality (Finland). The aliquots were thawed and labeled before distributed to participants the same day at ambient temperature. [d]Cobbaert C, Weykamp C, Franck P, de Jonge R, Kuypers A, Steigstra H, et al. Systematic monitoring of standardization and harmonization status with commutable EQA-samples – five year experience from the Netherlands. Clin Chim Acta 2012; 414:234-40 (PMID: 23041212). Samples were stored frozen at −70 °C, distributed on dry ice. [e]Blood was collected into dry blood bags by UK National Blood and Transplant Service at room temperature and allowed to clot at 4 °C. Serum was separated on the following day, then frozen at −40 °C and transferred frozen to UK NEQAS. Frozen serum was stored at −40 °C prior to thawing, pooling, aliquoting and refreezing at UK NEQAS. Specimens were distributed frozen and thawed in transit at ambient temperature. [f]https://www.niddk.nih.gov/health-information/communication-programs/nkdep/laboratory-evaluation/glomerular-filtration-rate/creatinine-standardization/commutability-study. Accessed 20 July 2020. [g]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. Available translated to English as a supplementary file in Clin Chim Acta 2012;414:234–40. [h]Referenzinstitut fur Bioanalytik, Cologne, Germany. [i]Reference Laboratory WEQAS, Cardiff, UK. [j]K=2.6 from t-distribution for 5 degrees of freedom.

data was first grouped according to method (enzymatic or Jaffe); and then by manufacturer, instrument platform and instrument model as shown in Supplementary Table 1. Unspecified information was marked as "other." Beckman and Siemens each had two distinct measurement procedures with different reagent formulations that were shown separately in the data aggregation. We found that the detail of the instrument specifications differed between EQA providers, consequently it was only possible to classify according to the least detailed category shown in Supplementary Table 1. For example, classification separately into Cobas and Modular measuring systems was not possible with the current data. Classification was also different or absent for some EQA providers. For example with Roche Cobas/Modular instruments, one EQA provider specified Cobas Instruments as 6000 or 8000, the second as c500 or c700, the third as 6000, 8000, c500 or c700 (or even c501/c702), while the fourth did not specify specific instruments (113 results, 36% of all Cobas/Modular results).

Calculations and statistical analysis

Outliers were excluded when exceeding ±3 SD of overall means per EQA provider. Bias vs. target value was calculated for each result from each EQA scheme. The mean % biases were derived from each individual bias. The expanded uncertainty of the mean bias for participants' results (U_{bias}) was calculated as two times the standard error of the mean (SEM) of the individual participants' % bias results (SEM = SD/$\sqrt{n}$). The combined expanded uncertainty ($U_{combined}$) was calculated from the expanded uncertainty of the target provided by each EQA provider (U_{target}) and the U_{bias} for each EQA as

$$U_{combined} = \sqrt{\left(U_{target}^{\ 2} + U_{bias}^{\ 2}\right)}$$

The largest U_{target} of the four EQA providers (1.88%) was used to calculate the overall combined uncertainty for aggregated EQA data. Mean biases and corresponding uncertainties were evaluated on the level of methods, instrument types and instrument measuring systems. All calculations were programmed and executed using R studio. Z-scores for deviations from zero bias were calculated from mean % bias and combined uncertainties (z-score = (0 − mean % bias)/$U_{combined}$). P-values were determined using the standard z-value to p-value conversion table based on a normal distribution and multiplied by two for a two-sided test [8, 9].

Results

Bias per EQA scheme

In total 1,011 results were submitted of which nine were excluded from analysis: six were regarded as outliers and

three were reported with no measuring system specified. Figure 1 compares the mean biases per EQA scheme for enzymatic assays by instrument platforms when at least 10 results were available among all schemes combined. For enzymatic assays, some significant differences were observed between the EQA schemes for Beckman AU and Siemens Advia measurement platforms, but overall results are consistent between the schemes. In Table 2 the number of results per instrument platform after harmonization of nomenclature is shown per EQA scheme. Jaffe methods are not discussed regarding instrument platforms as one EQA scheme provided only one result and another EQA scheme only 26 results (approximately four results per instrument platform).

Bias after data aggregation

After aggregation, mean % bias for the different methods is based on 567 and 435 numbers of results for enzymatic and Jaffe assays respectively, as presented in Figure 2A. These results demonstrate a statistically significant bias vs. reference measurement procedure (RMP) target values for Jaffe assays. When zooming in on enzymatic assays in Figure 2B, we observed that Abbott Architect and Siemens Advia demonstrated statistically significant negative biases of –2.7 and –8.2, respectively ($p < 0.01$) while Roche Cobas/Modular demonstrated a positive bias of 3.8% ($p < 0.01$). For Roche Cobas/Modular and Siemens Advia, this bias was also observed for the EQA results from each provider (Figure 1). As an example of examining individual instrument models, 314 enzymatic results from the Roche Cobas/Modular platform are shown in Figure 2C. Of note in Figure 2C are the 113 (36%) of participants who did not record the type of instrument model used.

We examined the influence of differences in mean % bias observed in Figure 1 for the UKNEQAS scheme vs. the other schemes for the Beckman AU and the Siemens Advia measurement procedures on the aggregated results. Figure 3 shows the mean % bias for the aggregated results including (panel A) and excluding (panel B) the UKNEQAS results. Removing the results from the UKNEQAS scheme changed the mean % bias for agregated results for the Beckman AU and the Siemens Advia but did not influence the mean % bias for the other platforms.

Discussion

Results aggregation

Data from EQA schemes gives valuable information regarding performance of IVD measurement procedures and is a useful tool to evaluate standardization or harmonization of results among these products, as well as their metrological traceability to higher order reference systems. In this pilot we aggregated EQA data from four EQA providers in different countries to examine the feasibility to provide useful information. The value to aggregate data is to be able to strengthen conclusions about the performance of specific IVD measurement procedures and to examine uniformity of metrological traceability in different regions and countries.

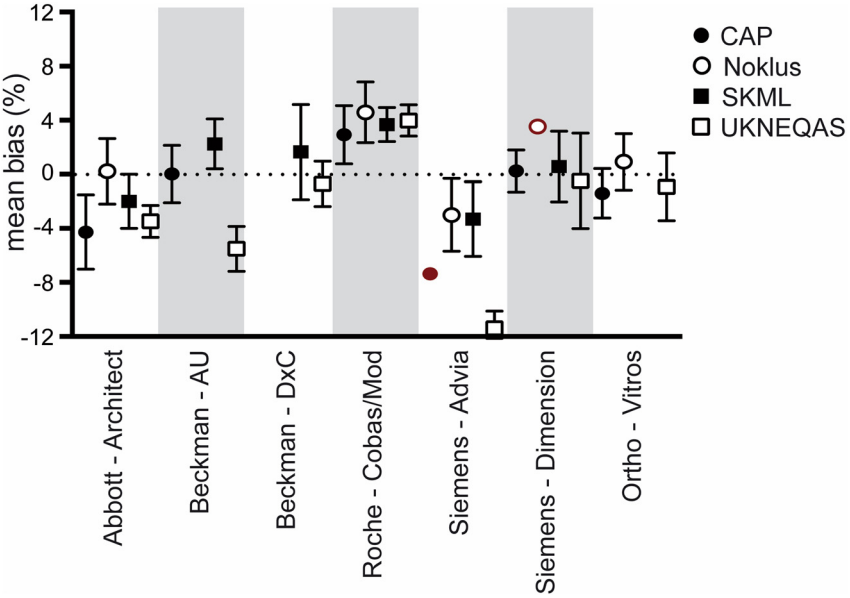

Figure 1: Creatinine results per EQA scheme for enzymatic methods for the major instrument platforms.
Mean % bias with combined expanded uncertainties are shown for the four EQA schemes. See Table 2 for numbers of results by manufacturer in each EQA scheme. Results were excluded when no measurement procedure was specified or when less than 10 results were available among all schemes combined. The red marker means that no uncertainty could be calculated (n=1).

Table 2: Number of results per Instrument type per EQAS for enzymatic methods.

Manufacturer	Instrument platform	CAP	NOKLUS	SKML	UKNEQAS	Total
Abbott	Architect	12	14	11	48	85
Beckman	AU	3	–	10	17	30
	DxC	–	–	8	2	10
Roche	Cobas/Modular	37	45	119	113	314
Siemens	Advia	1	6	9	26	42
	Dimension	5	1	11	3	20
Ortho	Vitros	35	5	–	8	48
Other		10	2	2	4	16
Total		103	73	170	221	567

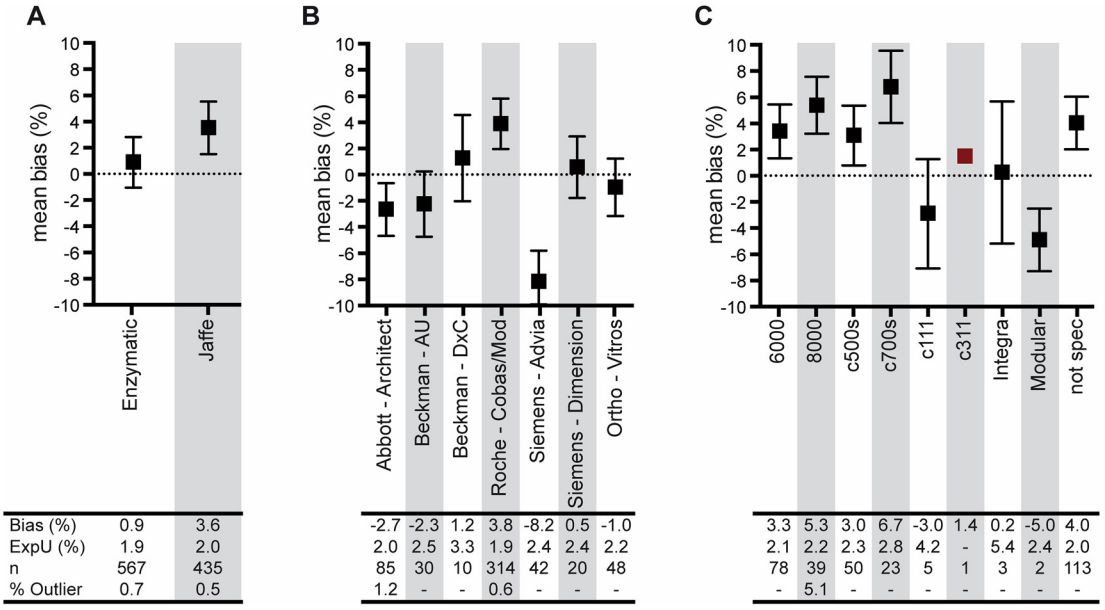

Figure 2: Creatinine results aggregated from four EQA providers.
Mean % bias with combined expanded uncertainties are shown for (A) the Jaffe and enzymatic methods, (B) results for enzymatic methods for the different instrument platforms with ≥10 instruments, (C) results for enzymatic methods for the different Roche Cobas/Modular instrument models. Below the graphs is shown the mean bias, combined expanded uncertainty, number of results and percentage of outliers. The red marker means that no uncertainty could be calculated (n=1).

We learned from the example data for creatinine that information on specific IVD measurement procedures can be obtained as well as more general information on instrument families and agreement of results among IVD manufacturers' products. For example, it is clear that enzymatic methods had a smaller bias (better trueness) than Jaffe methods for all IVD products (Figure 2A). Also, Siemens Advia enzymatic assays showed a significant negative bias that was observed in all four EQA schemes (Figure 1) as well as in the aggregated data (Figure 2B) with a mean bias of −8.2%. Importantly, data aggregation enabled larger numbers of results for each IVD measuring system to be examined. For example, aggregated results for the Roche Cobas/Modular group were based on measurements from 314 instruments (Figure 2B). Although, within the Roche Cobas/Modular group, the number of individual instrument models ranged from 1 to 78 and some had positive and some had negative biases (Figure 2C). Furthermore, for the Roche Cobas/Modular group, 113 of the 314 results (36%) were from an EQA scheme that did not identify the individual instrument models. For some IVD groups such as Beckman DxC only 10 aggregated results were available (Figure 2B) and for instruments such as Horiba ABX Pentra and Siemens Atellica only two results were submitted (Supplementary Table 1) that precludes useful conclusions for these groups. Aggregated information on specific IVD measuring systems is most useful for assessment of bias. Expanding the number of EQA schemes

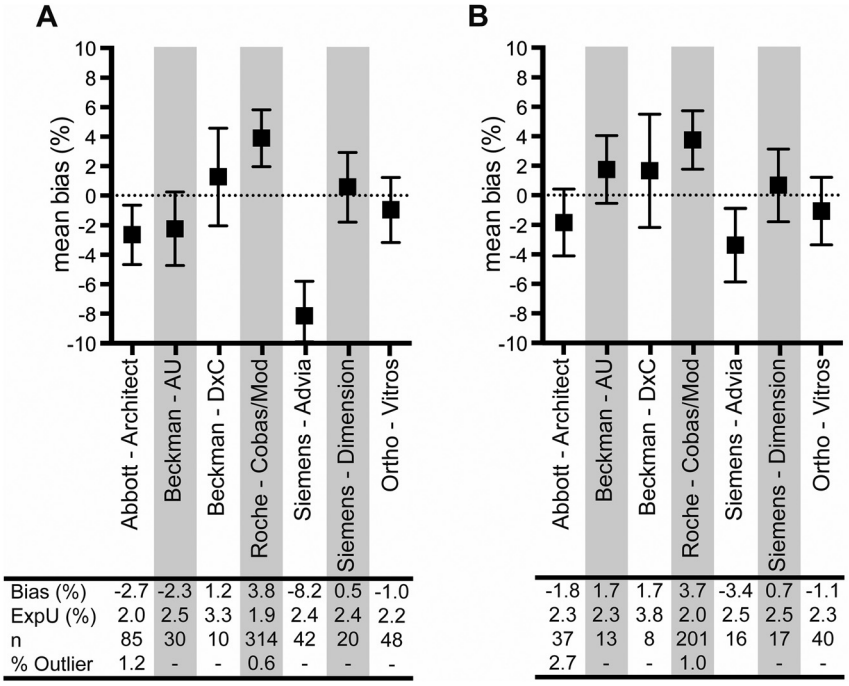

Figure 3: Mean % bias for the aggregated results including and excluding the UKNEQAS results.
Panel A (same as Figure 2B) shows the mean % bias for aggregated enzymatic creatinine results including results for the UKNEQAS scheme. Panel B shows the mean % bias after removing the results from the UKNEQAS scheme.

in the aggregation will achieve useful numbers of results from more IVD manufacturers' measuring systems.

Important points identified for extending the pilot to aggregate data from larger numbers of EQA providers include the following. First, data sets from individual EQA providers were successfully aggregated for statistical evaluation and presentation in a single format. Second, data in each EQA scheme's usual format had to be transformed to be able to classify method, manufacturer, instrument platform and instrument model into a common format. Going forward, dedicated software will be necessary to transform data from individual EQA scheme formats to a common format for analysis. Third, details of method, instrument platform and instrument model were not uniform among EQA schemes. Going forward, nomenclature and level of detail available from each EQA scheme needs to be standardized to describe methods, reagents, instrument platforms and instrument models using a common format to enable this data to be appropriately used to assess the status of metrological traceability and agreement among results.

A point not addressed in this pilot is how to determine what concentrations of EQA results are suitable to be aggregated to calculate a mean % bias. Since SD and CV can vary with concentration, criteria need to be developed to identify results from an interval of concentrations when either the SD or CV is sufficiently constant to support aggregation of results with an acceptable uncertainty in the mean % bias.

Commutability

A key requirement for data aggregation is that the EQA samples are commutable with clinical samples. Ideally, evidence of commutability should be provided by EQA providers. The pilot EQA schemes differed in documentation of commutability of the samples used. Some schemes had conducted formal commutability assessment for earlier batches of materials prepared according to the same process used for the current EQA samples. In these cases, an assumption was made that the current samples had the same commutability as earlier batches. Other schemes had no formal commutability assessment and assumed the samples were likely to be commutable based on how they were prepared. Figure 1 shows that there are significant differences especially between Beckman AU and Siemens Advia EQA results among the EQA providers. We cannot exclude that these differences are due to non-commutability of the EQA control material for some of the EQA providers. However, differences can be explained by calibration issues or calibrator or reagent lots used in the different countries. When differences between EQA schemes are observed, the data needs to be examined for suitability to be aggregated. For example, Figure 3 shows the results from the different platforms for each EQA scheme with and without the UKNEQAS results. Different conclusions for the Beckman AU and Siemens-Advia would be made; consequently how to present such data needs to

be considered when expanding the pilot to include more EQA schemes.

The differences observed in Figure 2B,C for the aggregated data, most likely reflect the different calibration status for the different platforms or different instrument types although non-commutability of the control material could also play a role. However, the preparation of each EQA sample was from minimally manipulated serum making non-commutability less probable since the uncertainty of the aggregated biases was reasonably small.

In any case, these findings emphasize that criteria for acceptable evidence of commutability for EQA samples needs to be developed for EQA providers to play a role in monitoring the status of harmonization and standarization efforts. An IFCC working group is currently addressing criteria for commutability of EQA as well as for certified reference materials and trueness controls. In addition, criteria need to be developed to qualify an EQA sample for inclusion in a data aggregation process. Commutability assessment is time consuming and expensive [10, 11]. Approaches for assessing commutability of EQA samples that are prepared frequently will probably be different than for certified reference materials that are expected to be stable for many years. Approaches for verifying commutability of replacement batches of EQA samples are also needed.

Target values for EQA samples

A target value for the EQA sample is required because the relative bias for each result to the assigned target value is the parameter used for data aggregation. Ideally, target values and uncertainties should be determined using a RMP listed by the Joint Committee for Traceability in Laboratory Medicine (JCTLM). Such target values enable assessment of metrological traceability of results to an established reference system. In this pilot, the target value for Noklus was established by a secondary value transfer process from a RMP value demonstrating the practical challenges faced by EQA providers and the necessity to apply pragmatic approaches, and use the data within the uncertainty limitations of the approach used. Going forward, an important qualification is how a target value was assigned because inconsistencies in the assigned value will influence the apparent biases observed for results from different EQA schemes. Trueness is defined as closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value for results [12]. Trueness is achieved by having metrological traceability to higher order references. Consequently, by aggregating the mean biases of results from different EQA schemes for a given measurement procedure, we are assessing the effectiveness of that procedure's metrological traceability to produce values with the property of trueness. For measurands for which RMPs are not available, harmonization of measurement procedures should be performed and how to set target values should be agreed. The uncertainty of a target value will be influenced by how it is derived and must be determined as part of the value assignment process.

Information describing measuring systems used by participants

This pilot identified that some EQA schemes did not collect sufficiently detailed information on instrument platform and the specific instrument model, measurement method and reagent to support aggregating the results. Differences in naming, e.g., Ortho Clinical Diagnostics vs. Vitros, are easily solved by renaming in the software data aggregation tool. For a data aggregation program to be optimal, measuring system details reported by participants in EQA schemes should be standardized among providers and a consensus agreed on the amount of detail to be included. In addition to the commonly reported instrument platform, method and reagent used, information on calibrator and reagent lots will improve the quality of information available for feedback to laboratories and to IVD manufacturers for evaluation of metrological traceability. Table 3 presents desirable information to be collected from participants to enable EQA data to be suitably aggregated among different EQA schemes. An additional component for consideration is to have a documentation and reporting system that participants can easily comply with to ensure the reported information is correct. A standardized set of parameters to classify measuring systems used by all EQA providers will promote participant compliance as well as ensure suitable information to aggregate data to monitor the status of harmonization and standardization of test results.

Limitations

A limitation of this pilot study was that results from only four EQA providers were available that limited the number of results in some groupings of instruments. In addition, the information available from some EQA providers was not sufficient to classify results by instrument models on the market from a given IVD manufacturer. For example, with the current data it was not possible to aggregate all of the data for

Table 3: Participant information needed for aggregation of results from different EQA providers.

Information	Minimum requirement	Desirable information	Example
Instrument manufacturer	X		Abbott
Instrument name	X		Architect
Instrument measuring system designation	X		C8000
Method type (reagent type)	X		Enzymatic
Reagent manufacturer		X	Abbott
Reagent lot number		X	R49872
Calibrator manufacturer		X	Abbott
Calibrator lot number		X	C43256
Calibration traceability (when applicable)		X	IDMS listed by JCTLM

Roche instruments separately into Cobas and Modular instruments, or into specific instrument models which would be desirable. The long term goal to aggregate data from a large number of EQA providers will provide sufficient individual results to enable suitable bias data to be reported for most instrument models in use in laboratory medicine.

Another limitation was that the commutability of all EQA samples was not rigorously assessed and assumptions had to be made based on the preparation procedures used. The biases observed in the aggregated data could therefore have been influenced by possible non-commutability of samples. One approach to examine this possibility is to remove data from one EQA scheme and determine if the conclusions are different. An indication of potential non-commutability can be assessed by comparing results from individual schemes to determine if different biases are observed. However, the assessment is complicated because the differences could be caused by non-commutability of the samples but could also be caused by country specific differences in calibration, reagent lots, small number of participants, etc. As stated previouisly, criteria need to be developed for acceptable commutability of EQA samples for which data will be aggregated.

Conclusions

Aggregation of EQA data from different providers can be accomplished and provides useful information regarding the biases among different measuring systems. A prerequisite for aggregating the data is that the EQA control material is commutable and that reporting the measurement methods, reagents, instrument platforms and models used by participants are standardized among EQA providers. Additional development is needed for criteria for suitable commutability of EQA materials for data to be aggregated, and to standardize the information collected from participants to optimize the classification of performance for different measurement procedures.

Research funding: None declared.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

References

1. ISO 17511. In vitro diagnostic medical devices — Requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples, 2nd ed. Geneva, Switzerland: International Organization for Standardization; 2020.
2. Miller WG, Jones GRD, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
3. Braga F, Pasqualetti S, Panteghini M. The role of external quality assessment in the verification of in vitro medical diagnostics in the traceability era. Clin Biochem 2018;57: 23–8.
4. Weykamp C, Secchiero S, Plebani M, Thelen M, Cobbaert C, Thomas A, et al. Analytical performance of 17 general chemistry analytes across countries and across manufacturers in the INPUtS project of EQA organizers in Italy, The Netherlands, Portugal, United Kingdom and Spain. Clin Chem Lab Med 2017; 55:203–11.
5. The EurA_{1c} Trial Group. The European HbA_{1c} trial to investigate the performance of HbA_{1c} assays in 2166 laboratories across 17 countries and 24 manufacturers using the IFCC Model for Quality Targets. Clin Chem 2018;64:1183–92.
6. Stöckl D, Reinauer H. Candidate reference methods for determining target values for cholesterol, creatinine, uric acid, and glucose in external quality assessment and internal accuracy control. I. Method setup. Clin Chem 1993;39:993–1000.
7. Thienpont LM, Leenheer AP, Stöckl D, Reinauer H. Candidate reference methods for determining target values for cholesterol, creatinine, uric acid, and glucose in external quality assessment and internal accuracy control. II. Method transfer. Clin Chem 1993; 39:1001–6.
8. Moore DS, McCabe GP, Craig BA. Introduction to the practice of statistics, 6th ed. New York: W. H. Freeman and Company; 2009.

9. P-value calculator, Graph Pad. Available from: https://www.graphpad.com/quickcalcs/pvalue1.cfm [Accessed 21 July 2020].
10. Miller WG, Schimmel H, Rej R, Greenberg N, Ceriotti F, Burns C, et al. For the IFCC working group on commutability. IFCC working group recommendations for assessing commutability Part 1: general experimental design. Clin Chem 2018;64:447–54.
11. Nilsson G, Budd JR, Greenberg N, Delatour V, Rej R, Panteghini M, et al. For the IFCC working group on commutability. IFCC working group recommendations for assessing commutability Part 2: using the difference in bias between a reference material and clinical samples. Clin Chem 2018;64:455–64.
12. International vocabulary of metrology – basic and general concepts and associated terms (VIM) JCGM, Joint Committee for Guides in Metrology, 3rd ed. vol. 200. JCGM; 2012.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2020-0736).