Lack of grading agreement among international hemostasis external quality assessment programs

John D. Olson[a], Ian Jennings[b], Piet Meijer[c], Chantal Bon[d], Roslyn Bonar[e], Emmanuel J. Favaloro[f], Russell A. Higgins[g], Michael Keeney[h], Joy Mammen[i], Richard A. Marlar[j], Roland Meley[k], Sukesh C. Nair[i], William L. Nichols[l], Anne Raby[h], Joan C. Reverter[m], Alok Srivastava[n] and Isobel Walker[b]

Laboratory quality programs rely on internal quality control and external quality assessment (EQA). EQA programs provide unknown specimens for the laboratory to test. The laboratory's result is compared with other (peer) laboratories performing the same test. EQA programs assign target values using a variety of methods statistical tools and performance assessment of 'pass' or 'fail' is made. EQA provider members of the international organization, external quality assurance in thrombosis and hemostasis, took part in a study to compare outcome of performance analysis using the same data set of laboratory results. Eleven EQA organizations using eight different analytical approaches participated. Data for a normal and prolonged activated partial thromboplastin time (aPTT) and a normal and reduced factor VIII (FVIII) from 218 laboratories were sent to the EQA providers who analyzed the data set using their method of evaluation for aPTT and FVIII, determining the performance for each laboratory record in the data set. Providers also summarized their statistical approach to assignment of target values and laboratory performance. Each laboratory record in the data set was graded pass/fail by all EQA providers for each of the four analytes. There was a lack of agreement of pass/fail grading among EQA programs. Discordance in the grading was 17.9 and 11% of normal and prolonged aPTT results, respectively, and 20.2 and 17.4% of normal and reduced FVIII results, respectively. All EQA programs in this study employed statistical methods compliant with the International Standardization Organization (ISO), ISO 13528, yet the evaluation of laboratory results for all four analytes showed remarkable grading discordance. *Blood Coagul Fibrinolysis* 29:111–119

Blood Coagulation and Fibrinolysis 2018, **29**:111–119

Keywords: coagulation, external quality assessment, hemostasis, proficiency testing

[a]Department of Pathology, University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA, [b]United Kingdom National External Quality Assurance Service - Blood Coagulation (UK-NEQAS-BC), Sheffield, UK, [c]ECAT Founation, Voorschoten, The Netherlands, [d]Association ProBioQual, Lyon, France, [e]RCPAQAP Haematology, [f]Department of Haematology, Institute of Clinical Pathology and Medical Research, Westmead Hospital, Sydney, New South Wales, Australia, [g]College of American Pathologists, Northfield, Illinois, USA, [h]Institute for Quality Management in Healthcare, Toronto, Ontario, Canada, [i]Department of Immunohematology and Transfusion Medicine, Christian Medical College, Vellore, Tamil Nadu, India, [j]Department of Pathology, University of New Mexico, Albuquerque, New Mexico, USA, [k]Laboratoire d'Hématologie, Hôpital Nord, Saint-Etienne, France, [l]Special Coagulation Laboratory (Hematopathology), Mayo Clinic, Rochester, Minnesota, USA, [m]Laboratorio de Evaluacion Externa de la Calidad en Hematologia (LEECH), Hospital Clinic Barcelona, Barcelona, Spain and [n]Department of Haematology, Christian Medical College, Vellore, Tamil Nadu, India

Correspondence to John D. Olson, MD, PhD, Professor Emeritus, Department of Pathology, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900, USA
E-mail: OLSONJ@uthscsa.edu

Received 31 May 2017 **Revised** 20 November 2017
Accepted 29 November 2017

Introduction

Laboratory quality programs rely heavily on both internal and external quality control [1]. Both of these methods depend on processes that help to assure that the laboratory method generates a reliable value to be reported. Internal quality control (IQC) takes advantage of testing specimens with 'known' or assigned values. Statistical control limits are then determined for each of the controls. Before a patient result can be reported, the control specimens must have results that fall within the limits that have been defined for each. If any control specimen value exceeds the defined limits, patient testing must be halted and the analytic method examined, with corrective action taken to bring the controls within their defined limits before reinstituting patient testing. Regulatory agencies may dictate the minimum required IQC practices for some analytes (e.g. number of controls used and the frequency that they must be run). This process controls and documents primarily the imprecision of the assay.

External quality assessment (EQA) is a second level of control for laboratory tests. EQA is a process by which an external program provides an unknown specimen for the laboratory to test. The result obtained by the laboratory and its interpretation are returned to the EQA program for comparison with other laboratories that perform the same test (their 'peers'). EQA programs then analyze the peer group data using a variety of statistical tools (including

0957-5235 Copyright © 2018 The Author(s). Published by Wolters Kluwer Health, Inc. DOI:10.1097/MBC.0000000000000690
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.bloodcoagulation.com).

comparison with all methods, comparison of peer groups using the same reagents, comparison of peer groups using the same instrument and reagent) to provide a target value (typically the median value or trimmed mean of the 'same' group) for the test with acceptable limits for the performance of a laboratory. The EQA providers also use variable methods to determine the acceptable range around the target value and will use different methods for different analytes. Methods include such techniques as parametric (e.g. ± 2 SD), nonparametric (e.g. $\pm 5\%$) or even visual inspection of the data by 'experts'. The EQA provider may provide a graded performance analysis (e.g. pass vs. fail), or the laboratory may assess their own performance relative to comparison data. In either case, the laboratory should take corrective action based on the results of the comparison if performance is beyond the limits set by the EQA program, particularly if there are persistent failing results on sequential surveys. This process documents the accuracy of the method.

In some countries, the EQA program must report laboratories with repeatedly unsatisfactory performance to the regulatory agencies, and punitive consequences may result, Center for Medicare and Medicaid Services through the Clinical Laboratory Improvement Act (CLIA) in the United States of America [2] and RiliBÄEK in Germany [3] being examples. When EQA is reported to an agency for licensure or accreditation the process is referred to as proficiency testing. Some regulatory agencies define the acceptable limits around the mean or target value for some tests (e.g. regulated analytes under CLIA). Even when regulatory agencies define the acceptable limits around mean or target value in policy or law, these agencies, at times, fail to recognize the differences in statistical methods used by EQA providers to arrive at mean or target values, potentially limiting the utility of regulatory oversight of proficiency testing. The majority of laboratories participate in only a single EQA program for each analyte that they report. Thus, comparison of the performance of the EQA program, or consistency among EQA programs, is difficult to assess. Although EQA programs should be accredited by a relevant accreditation body to international standards [4], there are many acceptable approaches to performance analysis for proficiency testing [5]. These published guidelines have been helpful to EQA programs, but still allow for considerable variability in the acceptable statistical methods used for evaluation.

Variability of performance analysis among EQA programs has been previously reported as have recommendations to standardized evaluation processes. The lack of harmonization among EQA programs may begin with the variability of the limits of acceptability that exist among programs [6–8]. Friedecky *et al.* reported wide variation in required limits of 16 clinical chemistry analytes across five national programs. Variation reported occurred in both the type and magnitude of units used [9]. In 2008, a study of the comparison of EQA programs grading of a data set of hemoglobin (Hb) and leukocyte count was reported. Among 11 programs, the failure rate for Hb ranged from 0.4 to 15.6% of laboratories, and for leukocyte count, 0–19.8% [10].

In 2005, an international group of EQA providers formed the external quality assurance in thrombosis and hemostasis (EQATH) group, with the goal of improving laboratory performance internationally through EQA by sharing information and ideas among EQA programs. By pooling knowledge, expertise and experiences, particularly regarding results of EQA challenges and sharing of ideas regarding program functions, it is hoped that improvements to individual programs and consistency among EQA programs may be achieved; by extension, laboratories internationally will benefit from such a collaboration. Fourteen EQA/proficiency testing programs are currently members of the EQATH group [11,12].

Anecdotally, two of the members of the EQATH group reported that their individual laboratories were participating in more than one EQA program for individual analytes. On most occasions, the results of the two EQA programs would be concordant, that is both programs would report a passing or failing grade. However, on more than one occasion, they had received a passing grade from one EQA program, but a failing grade from the other, for the same test or assay. These were, of course, different specimens, but the occurrences were close in time, and there was no clear explanation for the discrepant grading. When such discrepant or discordant results occur, the laboratory is faced with a dilemma as to what action if any should be taken.

As a result of the anecdotal reports from laboratories participating in more than one EQA and the previous literature reports, the EQATH group embarked on the study reported here. The goal of the study was to describe the degree of concordance/nonconcordance that may occur among EQA programs when analyzing the same data reported by a laboratory, particularly in respect of target result setting and passing/failing laboratory performance. To this end, a data set of results from multiple laboratories was compiled, and the same data set was sent to several EQA programs for analysis, and those analyses were then compared.

Methods

Participating external quality assessment programs

The programs that participated in this exercise are included in Appendix 1, http://links.lww.com/BCF/A42. World Heath Organization (WHO) and World Federation of Hemophilia (WFH) are administered by NEQAS, using the identical evaluation method – one single data set was returned for these three programs. Christian Medical College-External Quality Assurance Scheme (CMC-EQAS) uses a method similar to NEQAS.

North American Specialized Coagulation Laboratory Association (NASCOLA) is a part of the External quality Control of Diagnostic Assays and Controls (ECAT) program – one single data set was returned for these two programs.

Of these 14 programs participating in the EQATH collaboration, 11 programs from eight countries using eight different approaches to the statistical analysis participated in the data set analysis.

Data set

Data from a subset of the laboratory submissions to the NEQAS-BC program were extracted for the purpose of this study. Records were anonymized. Each record contained data for four EQA challenges returned by the laboratory: normal activated partial thromboplastin time (aPTT), prolonged aPTT, normal factor VIII (FVIII) assay and reduced FVIII assay. In addition, each record contained data to allow participant laboratories to be placed in method peer groups. The data elements included instrument, reagent, upper limit of the reference interval, lower limit of the reference interval, test result, test interpretation. For FVIII, additional data elements were the source of the reference plasma and the source of the deficient plasma. Records for 218 laboratories were included in the data set. For both the aPTT and the FVIII assay, the data set included up to 10 method peer groups (based on instrument and reagent, and the peer grouping approach used by EQA program) ranging in size from eight to 50 laboratories per group. To preserve anonymity, the EQA programs were randomly assigned numbers in the presentation of the data. One program did not provide analysis for the aPTT analyte; thus, only seven method approaches are described for the aPTT.

Statistical analysis

Each of the programs used a statistical method designed by the specific program for analysis of the data. Although similar, the methods for evaluation of the aPTT and FVIII data were not identical in all cases. The methods used are summarized below. The language used for acceptable performance is unique to each program. No program uses the terms 'pass' and 'fail'; however, for the purpose of this study, pass and fail are used for uniformity. A failing grade was assigned if the program grade would advise or require some action on the part of the laboratory.

Activated partial thromboplastin time statistical methods

The approach to analysis of aPTT data by each of the seven programs using unique methods of assessment is summarized in Table 1. The complexity of the analysis of some programs does not lend them to tabular presentation. Additional notes from those programs are as follows:

Program 1

Values of less than ±100% are considered passing. Values ±100–150% receive a warning with advice to investigate and values more than ±150% require action. For the purpose of this analysis, values greater than 100% were failing.

Program 6

Allowable limits of performance (ALP) are unique for each analyte and are calculated from the target overall median value of the instrument or reagent peer group with 10 or more users, whichever is applicable, and are used in the histograms and Youden plots. The limits are set based on clinical needs and are set and reviewed by program organizers and expert committee members.

Table 1 **Statistical methods used by programs for evaluation of the activated partial thromboplastin time**

Program	Units	Target	Peer group	Minimum number in peer group	Assessment method	Limits
1	Seconds	'Adjusted' mean from median and iterative process	Agent & instrument	Min 5–8	% Deviation from peer group median	±25% deviation[a]
2	Ratio	Median	Reagent (overall if $n < 10$)	10	% Deviation	±15%
3	Seconds	Mean following $2x \pm 3$ SD passes (+Dixons Q test to exclude outliers)	Reagent & instrument	10	% Deviation	±15%
4	Seconds	Mean following $2x \pm 3$ SD passes	Reagent & instrument	Not stated	SD index	±2.0
5	Seconds	Median/truncated mean after 1 pass of ±3 SD[b]	Overall	–	% Deviation	±15%
6	Seconds	Median	Reagent/instrument as appropriate	10	Acceptable performance limits, set by committee	±10 up to 40 s; ±25% >40 s
7	Seconds	Truncated mean after 2 passes of ±2 SD	Reagent & instrument (also overall)	Min 5–8	% Deviation	Adjusted for a number of factors: % deviation from mean and *z*-score (see text)

[a] PAD = $[(x - x_a)/\text{APL}] \times 100$; values $< \pm 100$ pass. Where PAD is percentage allowable difference, x is the participant result, x_a is the 'Assigned Value' and APL is the Allowable Performance Limit (see text). [b] Mean if data are normally distributed, median if not normally distributed.

Table 2 Statistical methods used by programs for evaluation of the factor VIII assay

Program	Target	Peer group	Number in peer group	Assessment method	Limits
1	'Adjusted' mean from median and iterative process	Agent & instrument	Min 5	% Deviation from peer group median	±25% deviation
2	Median	Overall	–	A–E grading (see text)	A–C grades (see text)
3	Mean following $2x \pm 3$ SD passes (+Dixons Q test to exclude outliers)	Reagent & instrument; overall if no peer group	–	% From the mean	50% Criteria if mean <50 and 20% if mean >50
4	Mean following $2x \pm 3$ SD passes	Overall	–	Deviation index	±2.0
5	Median	Overall	–	A–E grading (see text)	A–C grades (see text)
6	Median	Overall/reagent/instrument as appropriate	10	Acceptable performance limits, set by committee	±3.0 up to 10 s; ±30% >10%
7	Truncated mean after 2 pass of ±2 SD	Deficient plasma and cephaline reagent (also overall)	Min 5–8	% Deviation	Adjusted for a number of factors: % of mean and additionally *z*-score <3 (see text)
8	Trimmed mean of data	Overall	–	*z*-Score	<3

Program 7

There are two grading systems used. The first uses acceptable limits of deviation from the mean using both the all method and peer group analysis. Passing performance must fall within a defined percentage of the mean (12% for normal aPTT and 15% for prolonged aPTT). Failing performance is graded ±1–5 (five being the furthest from the mean). If there is no target value by technique group (insufficient number of values), the result can only be noted in relation to the all methods group. The peer group method is preferred and given the highest consideration when grading.

The second grading system involves calculation of the *z*-score. A *z*-score more than 3 is failing, a *z*-score 2–3 is borderline or warning level and a *z*-score less than 2 is passing. Although most laboratories receiving a failing grade fail in both grading categories, there are occasionally laboratories that will pass one grading method, but fail the other grading method. In the case of this study, if a failing grade is received in either category, the laboratory is considered to have failed the exercise.

Factor VIII assay statistical methods

The approach to analysis of FVIII assay data by each of the eight programs using unique methods of assessment is summarized in Table 2. The complexity of the analysis of some programs does not lend them to tabular presentation. Additional notes from those programs are as follows:

Program 1

Allowable performance for FVIII is ±25% deviation limits based on all methods. However, grading is based on allowable difference (%, Table 2). Values of less than ±100% are considered passing. Values ±100–150% receive a warning with advice to investigate and values more than ±150% require action. For the purpose of this analysis, values greater than 100% were considered to be failing.

Programs 2 and 5

The programs evaluate an all method peer group. The central reference point is taken as the overall consensus median. Individual results are ranked into five unequal groups above and below the median, each group being designated by a letter depending on ranked distance from the median, with lower case letters (e.g. 'b') denoting a result that is below the median, and an upper case letter (e.g. 'B') denoting a result that is higher than the median. Grades reflecting the percentage of results between 'A' and 'E' are assigned: A – ±25%, B – ±26–35%, C – ±36–40%, D – ±41–45%, E – ±46–50%. A failing performance designation is based on grades obtained in *two consecutive* exercises for any particular test, with persistent failing performance defined as two consecutive failures. For the purposes of this analysis, values of Dd or Ee were considered failing.

Program 6

The ALP are based on clinical needs and are set and reviewed by program organizers and expert committee members.

Program 7

For FVIII by activated cephaline time (aPTT), the statistical calculations are carried out according to the deficient plasma and cephaline reagents. There are two grading systems used. The first uses acceptable limits of deviation from the mean using both the all method and peer group analysis. Passing performance must fall within a defined percentage of the mean (20% for normal FVIII and 30% for reduced FVIII). Failing performance is graded ±1–5 (five being the furthest from the mean). If there is no target value by technique group (insufficient number of values), the result can only be noted in relation to the all methods group. The peer group method is preferred and given the highest consideration when grading. The second grading system involves calculation of the *z*-score. A *z*-score more than 3 would be considered failing, *z*-score 2–3 is borderline or warning level and *z*-

Table 3 Activated partial thromboplastin time analysis for sample 1, prolonged activated partial thromboplastin time, showing those laboratories that demonstrate nonconcordance in the grading

Program number Laboratory number	1	2	3	4	5	6	7	Total no. of programs failing this center
19				Fail				1
37		Fail						1
55		Fail						1
59				Fail				1
114				Fail				1
133		Fail						1
140		Fail						1
157				Fail				1
159				Fail				1
177			Fail					1
199				Fail				1
206				Fail				1
211			Fail					1
72				Fail			Fail	2
148			Fail				Fail	2
168				Fail			Fail	2
39		Fail			Fail			2
164			Fail	Fail				2
173		Fail			Fail			2
181		Fail			Fail			2
194		Fail			Fail			2
196		Fail			Fail			2
209		Fail			Fail			2
132		Fail	Fail				Fail	3
172		Fail	Fail				Fail	3
169		Fail		Fail	Fail			3
41			Fail	Fail	Fail		Fail	4
47		Fail	Fail	Fail			Fail	4
80		Fail	Fail	Fail			Fail	4
85		Fail	Fail		Fail		Fail	4
153		Fail	Fail	Fail			Fail	4
184		Fail	Fail		Fail		Fail	4
136		Fail	Fail	Fail	Fail			4
27		Fail	Fail		Fail	Fail	Fail	5
195		Fail	Fail	Fail	Fail		Fail	5
128	Fail	Fail	Fail		Fail	Fail	Fail	6
139	Fail	Fail	Fail		Fail	Fail	Fail	6
142	Fail	Fail	Fail		Fail	Fail	Fail	6
163	Fail	Fail	Fail		Fail	Fail	Fail	6
7	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
134	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
147	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
154	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
201	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
Total numbers of centers passed	209	188	194	196	196	208	196	
Total numbers of centers failed	9	30	24	22	22	10	22	

score less than 2 is passing. Although most laboratories receiving a failing grade, fail in both grading categories, there are occasionally laboratories that will pass one grading method, but fail the other grading method. For the purpose of this study, if a failing grade is received in either category, the laboratory is considered to have failed the exercise.

Results

Activated partial thromboplastin time analysis

Table 3 shows the centers failing analysis of the prolonged aPTT sample, and Table 4 shows the centers failing the normal aPTT sample analysis. In both cases, the tables show the programs which failed each laboratory, and the total number of programs failing that laboratory. A summary of the concordance of the aPTT grading among programs as a function of the size of the instrument and/or reagent peer groups is presented in Table 5. The data in this table are compiled from the grades from six programs. Program 8 did not provide EQA for aPTT at the time of this study, and did not do the analysis. Program 5 used an all method approach to grading, not using peer groups for analysis and is not included in the peer group analysis.

Normal activated partial thromboplastin time

The all method mean value was 32.8 s and the median was 32.9 s. Of the 218 records analyzed, 174 (79.8%) records were graded passing by all programs, and five records (2.3%) were graded as failing by all programs. There were 39 (17.9%) records that had discordant grades (passed by some programs and failed by others).

Prolonged activated partial thromboplastin time

The all method mean value was 53.1 s and the median was 51.8 s. Of the 218 records analyzed, 193 (88.5%)

Table 4 Activated partial thromboplastin time analysis for sample 2, normal activated partial thromboplastin time, showing those laboratories that demonstrate nonconcordance in the grading

Program number Laboratory number	1	2	3	4	5	6	7	Total fails
159		Fail						1
168				Fail				1
194		Fail						1
199				Fail				1
49		Fail						1
102		Fail						1
104		Fail						1
125		Fail						1
133				Fail				1
27				Fail				1
184				Fail				1
157			Fail					1
177				Fail				1
195			Fail					1
211			Fail					1
1				Fail				1
63				Fail				1
38				Fail				1
41				Fail				1
207	Fail			Fail				2
32	Fail	Fail	Fail	Fail			Fail	5
134	Fail		Fail	Fail	Fail	Fail	Fail	6
147	Fail		Fail	Fail	Fail	Fail	Fail	6
154	Fail		Fail	Fail	Fail	Fail	Fail	6
201	Fail	Fail	Fail	Fail	Fail	Fail	Fail	7
Total numbers of centers passed	212	210	210	202	214	214	213	
Total numbers of centers failed	6	8	8	16	4	4	5	

records were graded passing by all programs and one record (0.5%) graded as failing by all programs. There were 24 (11%) records that had discordant grades.

Normal factor VIII

The all method mean value was 59.7 IU/dl, and the median was 59 IU/dl. The variation among the target values and acceptable ranges assigned or calculated for each program was less than 1 IU/dl for both the normal and reduced FVIII samples. Of the 218 records analyzed, 172 (78.9%) records were graded passing by all programs and two records (0.9%) graded as failing by all programs. There were 44 (20.2%) records that had discordant grades. The distribution of the grades determined by each program for the normal FVIII data set is shown in Table 6. Concordance among the program grading and the value of FVIII reported is presented in the histogram in Fig. 1.

Reduced factor VIII

The all method mean value was 23.8 IU/dl, and the median was 22 IU/dl. The variation among the target values and acceptable ranges assigned or calculated for each program was less than 1 IU/dl for both the normal and reduced FVIII samples. Of the 218 records analyzed, 174 (79.8%) records were graded passing by all programs and six records (2.8%) were graded failing by all programs. There were 38 (17.4%) records that had discordant grades. The distribution of the grades determined by each program for the reduced FVIII data set is shown in Table 6. This table shows a variable number of centers failing the normal FVIII sample challenge and the reduced FVIII sample challenge in each program anaylsis. Concordance among the program grading and the value of FVIII reported in presented in the histogram in Fig. 2.

Among the four sets of data analyzed, two aPTT and two FVIII assay, there were no two EQA programs that graded all laboratories in the data set the same.

Discussion

Laboratories participate in EQA programs to verify the accuracy of the method being used for patient testing. In addition, in some locations, success in proficiency testing is a requirement to maintain accreditation or licensure. Accrediting agencies as well as laboratories rely on the data that are generated from the EQA programs to reliably identify any methods used by the laboratory that do not meet a standard of performance. EQA programs are expected to seek accreditation with an appropriate body to a relevant international standard and to employ statistical methods for analysis compliant with international standards, such as ISO 13528. The EQA programs

Table 5 Concordance among program grading as a function of peer group size for the activated partial thromboplastin time

	Normal aPTT sample			Prolonged aPTT sample		
Peer group size (laboratories in each peer group)	Passed by all programs, *n* (%)	Failed by all programs, *n* (%)	[a]Nonconcordant (discordant), *n* (%)	Passed by all programs, *n* (%)	Failed by all programs, *n* (%)	[a]Nonconcordant (discordant), *n* (%)
8	7 (87.5)	0	1 (12.5)	7 (87.5)	0	1 (12.5)
11	4 (34.4)	3 (27.3)	4 (27.3)	4 (27.3)	0	7 (72.7)
18	18 (100)	0	0	18 (100)	0	0
19	15 (79.0)	0	4 (21.0)	18 (94.7)	0	1 (5.3)
19	16 (84.2)	0	3 (17.8)	17 (89.5)	0	2 (10.5)
21	21 (100)	0	0	21 (100)	0	0
23	12 (47.8)	0	11 (52.2)	20 (87.0)	0	3 (13.0)
24	17 (70.8)	0	7 (19.2)	20 (83.3)	0	4 (16.7)
25	19 (76.0)	2 (8.0)	4 (16.0)	21 (84.0)	1 (4.0)	3 (12.0)
50	46 (92.0)	0	4 (8.0)	47 (94.0)	0	3 (6.0)
Total – 218	174 (79.8)	5 (2.3)	39 (17.9)	193 (88.5)	1 (0.5)	24 (11.0)

aPTT, activated partial thromboplastin time. [a]Nonconcordance (discordant): a laboratory that is failed by one or more, but not all, programs.

Table 6 Distribution of grading of the factor VIII samples: data by program

	Program 1	Program 2	Program 3	Program 4	Program 5	Program 6	Program 7	Program 8
Normal FVIII sample								
Pass, *n*	213	178	189	205	213	208	190	215
Fail, *n*	5	40	29	13	5	10	28	3
Reduced FVIII sample								
Pass, *n*	203	180	209	200	201	195	200	214
Fail, *n*	15	38	9	18	17	23	18	4

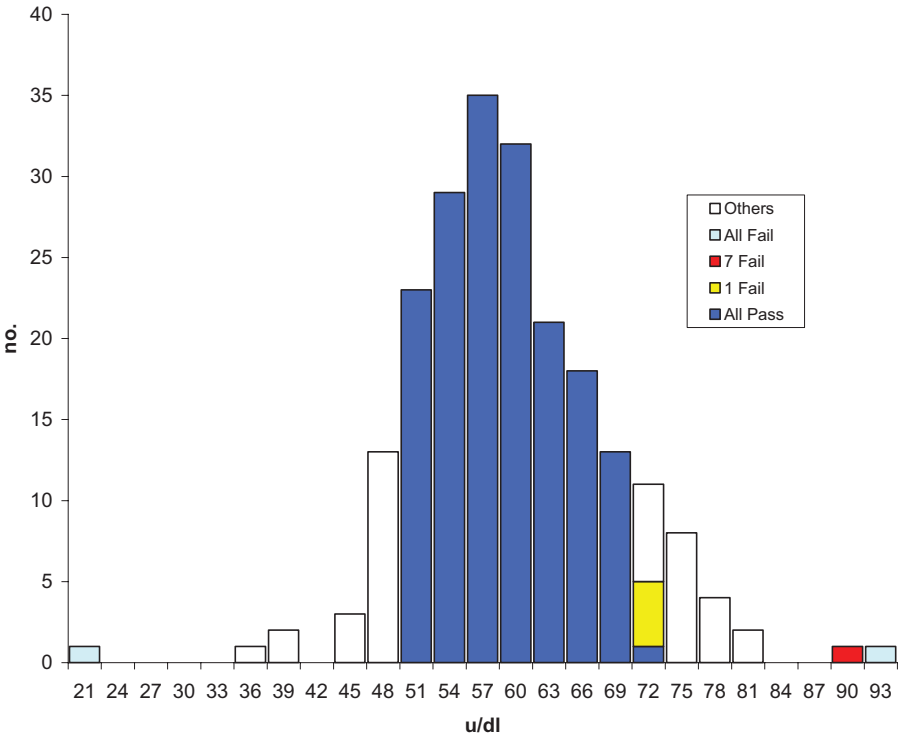
FVIII, factor VIII.

participating in this study used a variety of statistical methods to evaluate the data, some parametric and some nonparametric.

All programs employ consensus target values – either a median or trimmed mean, compatible with the ISO 13528 recommendations for value assignment, and all programs employ performance analyses including % deviation, *z*-scores or ranked grading analysis all of which are also compatible with the ISO guidelines [5]. Some programs used a different method for evaluation of the aPTT data than used to evaluate the FVIII data. Other differences in approaches included definition of peer groups, which in some cases were by reagent only and in others by reagent/instrument combination. One program assesses aPTT results in the form of ratios, whereas the others assess raw data in seconds. Two programs employ a ranked grading analysis for assays and assign performance based on grades achieved over two or more challenges – assignment of a 'fail' grade in these cases, based on grades awarded for a single challenge, was an artificial approach employed for this study only.

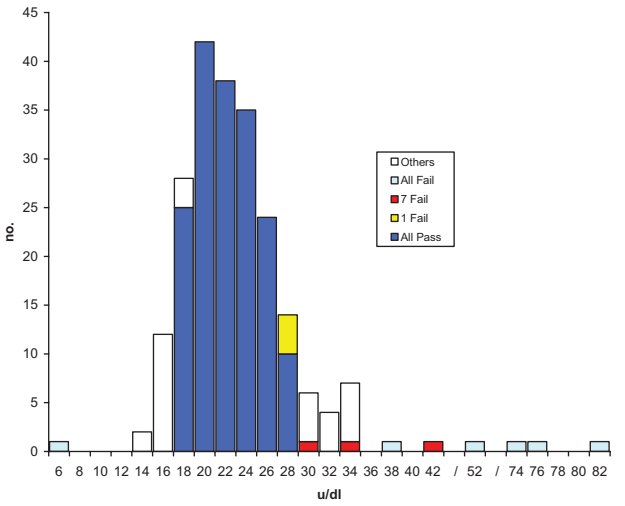
The results of the data analysis in this study demonstrate that pass/fail performance is variable among EQA programs. Perhaps inevitably, as the EQA programs use different methods for the analysis of the data, there is variability in the number of laboratories that are given passing or failing grades. Guidance for statistical evaluation of proficiency testing data exists in an international standard (ISO 13528: 2015). However, this standard describes a number of different approaches to evaluation

Fig. 1



Normal factor VIII sample: histogram of the laboratory grading as a function of the concentration of factor VIII reported (mean 59.7 IU/dl; median 59 IU/dl). Concordance of laboratories receiving a passing grade from all programs is royal blue and of all laboratories receiving a failing grade is light blue. Discordant records are represented as follows: failed by one program, passed by all the others – yellow; seven failing with others passing – red; between two and four programs failing – white.

Fig. 2



Reduced factor VIII sample: histogram of the laboratory grading as a function of the concentration of factor VIII reported (mean 24 IU/dl; median 22 IU/dl). Concordance of laboratories receiving a passing grade from all programs is royal blue and of all laboratories receiving a failing grade is light blue. Discordant records are represented as follows: failed by one program, passed by all the others – yellow; seven failing with others passing – red; between two and four programs failing – white.

of performance, and all the EQATH centers herein reported, used statistical approaches described in this standard.

From the data analysis described here, most laboratories were considered to have performed well, and there was good concordance among the programs for laboratories deemed as passing the challenges – ranging from a low of 78.9% of centers passed by all programs for the normal FVIII sample to a high of 88.5% passed by all programs for the prolonged aPTT sample. In contrast, agreement among the programs regarding the laboratories graded as failing showed very little concordance, with just one center considered to have failed the prolonged aPTT exercise by all programs and two centers deemed failing the normal FVIII assay. There were a substantial number of discordant records ranging from a low of 11% discordant records (in which grading from at least one program was at odds with the others) for the prolonged aPTT sample to a high of 20.2% discordant records for the normal FVIII sample. The increase in the number of discordant specimens for the FVIII samples is due to both the increased variance in the results reported by the 'laboratories and the differences among the statistical methods used by the programs. Some inconsistency was also apparent, for example in which a laboratory failed their FVIII assay, despite their result being closer to the median than another result that passed. These data are similar to the variability in failure rates reported for EQA programs grading Hb and leukocyte count [11].

Efforts to improve harmonization of the pass/fail assignment may be achievable. An encouraging finding in this study was that the target means or medians assigned for the four data sets were very similar. However, the approach to assigning acceptability limits around this value was highly variable and contributed to discordant pass/fail assignments among EQA providers. It would be possible for EQA providers to implement more uniform acceptability limits and improve harmonization of the performance analysis. A limitation to such approach includes instances in which acceptability limits for a test are mandated by a regulatory body. In the United States of America, the CLIA includes a list of 83 'regulated' analytes (aPTT is a regulated analyte) with prescribed acceptability limits. Nonetheless, there are many tests without such restrictions, and such prescribed acceptability limits could be the starting point for discussion. Notably, none of the EQA providers reported the use of an absolute value (e.g. $\pm$1 IU/dl for FVIII), an alternate way to express limits at very low levels of FVIII. This study highlights variability and provides an initial groundwork for future discussions regarding nonconformance among EQA providers. Laboratories participating in EQA programs rely on evaluation by the EQA provider to provide information to confirm the accuracy of their testing. When the EQA program gives a passing or failing grade that is incorrect, or inappropriate grading is applied, laboratories may be led to believe that a poorly performing test is satisfactory or that a test actually performing well is in need of corrective action. Incorrect assignment of a failing grade can also have negative impacts on licensure or accreditation when reporting to an agency. Each of the EQA programs participating in this study used a statistical method designed to fairly assess laboratory performance, in accordance with international standards and described by ISO, yet a significant number of laboratories had discordant scoring among the programs, highlighting the potential consequences of the different criteria employed.

The current study is intended to be descriptive of the methods used for EQA evaluation and the variability that is imparted to grading individual laboratory performance among the methods. It would be useful if the analysis would allow determination of the optimal scoring method among those used, but there are limitations that preclude that in this study. The first is that to determine the optimal scoring, one would need to know the 'truth', which of the laboratories are best performers and which are worst. Knowing this, evaluation of the ability of a scoring method to properly identify the truly best and worst laboratory performance could be evaluated. An additional limitation is that among the EQA programs, there may not be a single best method. The use of a nonparametric method demands that peer groups be

large and would only be optimal for programs with large peer groups. Alternately, the use of a parametric method demands that the data for a peer group be normally distributed, a condition that is not always met. It is entirely possible that a single evaluation method is not applicable to all analytes, and that an EQA program may need to use different scoring methods for different analytes as is done by some participants in this study.

Among the factors influencing the selection of statistical evaluation applied to EQA results an important one is that of commutability of the specimen. As a result of the various steps involved in creating EQA samples to desired specifications, the measurand and matrix undergo modifications that affect the commutability [13]. Until EQA organizers are able to provide cost effective, commutable samples, they will continue to evaluate participants on the basis of peer groups. This study also indicates that it is essential to harmonize analytic performance specifications. Ongoing efforts in the global laboratory community in building consensus as to what may be desirable limits based on the purpose of the test for example diagnostic vs. monitoring, the effect of analytical performance on clinical outcomes, components of biological variation of the measurand and on state-of-the-art which depends on the highest (currently) available level of analytic performance, respectively [14]. Further studies are required to test these premises especially in the specialty of hemostasis and offer evidence to validate the most appropriate model to be followed.

The information reported here indicates that there is a need for EQA programs to develop a more uniform approach to EQA evaluation and, at least, underlines the importance of utilizing statistically valid and clinically relevant performance criteria. Perhaps most importantly, there is an obligation amongst the professionals performing laboratory tests to evaluate the clinical as well as statistical relevance of their EQA performance.

Acknowledgements

The authors wish to thank the following individuals for their contribution to the acquisition of the data, the analysis of the data and assistance with the preparation of the article: Dr Thomas Long of the College of American Pathologists, Northfield, IL, USA; Tim Woods, Steve Kitchen and Diane Kitchen of UK NEQAS (Blood Coagulation), Sheffield, UK; Gabriela Gutierrez of Laboratorio de Evaluacion Externa de la Calidad en Hematologia (LEECH), Hospital Clínic Barcelona, Barcelona, Spain; John Soufi of RCPAQAP Haematology, St Leonards, NSW, Australia; Rita Selby and Karen Moffat of IQMH, Toronto, ON, Canada.

Author disclosures: I.J. and I.W. are compensated by UK-NEQAS; P.M. is compensated by ECAT; R.B. is compensated by RCPA-QAP; in the past, A.R. and M.K. were compensated by IQMH; other authors do not have disclosures regarding this work.

Conflicts of interest

There are no conflicts of interest.

References

1 Bonar R, Favaloro EJ, Adcock D. Quality in coagulation and haemostasis testing. *Biochemia Medica* 2010; **20**:184–199.
2 Department of Health and Human Services, Healthcare Financing Administration. Clinical laboratory improvement amendments of 1988; final rule. *Fed Regist* 1992; **57**:7001–7288.
3 Executive Board of the German Medical Association. Revision of the Guideline of the German Medical Association on Quality Assurance in Medical Laboratory Examinations – Rili-BÄEK. *J Lab Med* 2015; **39**: 26–69.
4 ISO/IEC 17043. Proficiency testing scheme providers ISO/IEC 17043 standard application document. 2014.
5 ISO 13528, statistical methods for use in proficiency testing by interlaboratory comparisons, 2nd ed. 2015.
6 Westgard JO, Seehafer JJ, Barry PL. European specifications for imprecision and inaccuracy compared with operating specifications that assure the quality required by US CLIA proficiency-testing criteria. *Clin Chem* 1994; **40 (7 Pt 1)**:1228–1232.
7 Sciacovelli L, Zardo L, Secchiero S, Plebani M. Quality specifications in EQA schemes: from theory to practice. *Clin Chim Acta* 2004; **346**:87–97.
8 Coucke W, China B, Delattre I, Lenga Y, Van Blerk M, Van Campenhout C, *et al.* Comparison of different approaches to evaluate external quality assessment data. *Clin Chim Acta* 2012; **413**:582–586.
9 Friedecky B, Kratochvila J, Budina M. Why do different EQA schemes have apparently different limits of acceptability? *Clin Chem Lab Med* 2011; **49**:743–745.
10 Van Blerk M, Albaré de S, Deom A, Gutiérrez G, Heller S, Nazor A, *et al.* Comparison of evaluation procedures used by European external quality assessment scheme organizers for haemoglobin concentration and leukocyte concentration. *Accred Qual Assur* 2008; **13**:145–148.
11 Olson JD, Preston FE, Nichols WL. External quality assurance in thrombosis and hemostasis: an international perspective. *Semin Thromb Hemost* 2007; **33**:220–225.
12 EQATH: http://eqath.org. Accessed 1 Dec 2017.
13 Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, *et al.* Defining analytical performance specifications: consensus statement from the 1st strategic conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015; **53**:833–835.
14 Miller WG, Myers GL, Rej R. Why commutability matters. *Clin Chem* 2006; **52**:553–554.