Piet Meijer*, Frederic Sobas and Panagiotis Tsiamyrtzis

Assessment of accuracy of laboratory testing results, relative to peer group consensus values in external quality control, by bivariate z-score analysis: the example of D-Dimer

https://doi.org/10.1515/cclm-2023-0835
Received August 2, 2023; accepted February 16, 2024;
published online March 11, 2024

Abstract

Objectives: The aim of this study is to develop a practical method for bivariate z-score analysis which can be applied to the survey of an external quality assessment programme.
Methods: To develop the bivariate z-score analysis, the results of four surveys of the international D-Dimer external quality assessment programme of 2022 of the ECAT Foundation were used. The proposed methodology starts by identifying the bivariate outliers, using a Supervised Sequential Hotelling T^2 control chart. The outlying data are removed, and all the remaining data are used to provide robust estimates of the parameters of the assumed underlying bivariate normal distribution. Based on these estimates two nested homocentric ellipses are drawn, corresponding to confidence levels of 95 and 99.7 %. The bivariate z-score plot described provides the laboratory with an indication of both systematic and random deviations from zero z-score values. The bivariate z-score analysis was examined within survey 2022-D4 across the three most frequently used methods.
Results: The number of z-score pairs included varied between 830 and 857 and the number of bivariate outliers varied between 20 and 28. The correlation between the z-score pairs varied between 0.431 and 0.647. The correlation between the z-score pairs for the three most frequently used varied between 0.208 and 0.636.
Conclusions: The use of the bivariate z-score analysis is of major importance when multiple samples are distributed around in the same survey and dependency of the results is likely. Important lessons can be drawn from the shape of the ellipse with respect to random and systematic deviations, while individual laboratories have been informed about their position in the state-of-the-art distribution and whether they have to deal with systematic and/or random deviations.

Keywords: external quality assessment; z-score; bivariate analysis; systematic error; random error

Introduction

It has been demonstrated that laboratory testing results play a role in up to 80 % of clinical decision making [1]. It has also been shown that laboratory testing can be included in up to 94 % of the recommendations of clinical guidelines [2]. An accurate diagnosis, frequently based on test results from the laboratory, is a prerequisite for adequate decisions on treatment [3]. On the other hand, diagnostic errors can count for 25–75 % of identified medical errors [4]. This emphasises the need for accurate test results. However, a laboratory test can be subject to systematic errors and random errors. It is therefore important that a clinical laboratory has implemented appropriate procedures for quality control, which includes both internal quality control and external quality assessment (EQA). The primary aim of EQA is to focus on the laboratory's analytical performance in comparison to the values assigned on the basis of an accuracy-based reference system or on peer-group consensus values [5]. Therefore, the major focus of analytical performance assessment by an EQA programme is the assessment of the deviation (bias) from the intended value. Several possibilities are described to assess the (relative) accuracy of measurement of a participant in an EQA programme [6]. A frequently used measure for the deviation between a participant's result and the assigned value is the z-score, which reflects this deviation corrected by the variation in the distribution of all participants' results [7, 8]. The z-score is especially helpful for comparing participants' performance of peer groups

***Corresponding author: Piet Meijer**, ECAT Foundation, Voorschoten, The Netherlands, E-mail: P.Meijer@ecat.nl. https://orcid.org/0000-0003-4899-3294
Frederic Sobas, Haemostasis Department, Hospices Civils de Lyon, Lyon, France
Panagiotis Tsiamyrtzis, Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy; and Department of Statistics, Athens University of Economics and Business, Athens, Greece

with consensus values which are not standardised nor harmonised. Z-scores can be evaluated separately for each sample included in a survey (univariate approach). When multiple samples are distributed in one survey the z-score pairs of all participants can be plotted in a Youden plot [8]. Acceptance limits are frequently plotted by rectangular areas, for instance at the level of z-score is –2 and 2. However, it has been demonstrated that this approach has its limitations [9]. In the case of a survey with multiple samples it is plausible that all samples are measured in the same analytical run and therefore are designated as dependent results. Z-score pairs should therefore be evaluated as paired measurements by a bivariate control chart [9–12]. In monitoring z-score pairs the naïve approach is to consider two individual control charts which aim to examine each z-score independently of the other, an approach that is not only inefficient but can be misleading [13]. The problem arises from the fact that while the data have a bivariate origin, their projection down to one variable at a time leads to loss of significant spatial information. Thus, a method that will allow us to monitor the pair of z-scores simultaneously, i.e. as a single entity, is in order. Specifically, a bivariate approach, which will take into account the correlation between the two z-scores is needed. In general, the first ever multivariate control chart approach initiated by Hotelling [14] can be used, but since Hotelling, a large variety of methods has been developed [15]. The application of a multivariate approach to quality control results in the area of the clinical laboratory has been described previously as a useful tool to detect shifts in internal quality control results [16].

The D-Dimer assay results are typically of great importance for clinical decision making in haemostasis. The D-Dimer serves as a valuable marker of the activation of coagulation and fibrinolysis in many clinical scenarios [17]. Most commonly, D-Dimer has been extensively investigated for excluding the diagnosis of deep venous thrombosis (DVT) [18]. Using the cut-off value advised by the manufacturer, the D-Dimer test is useful to exclude DVT safely, omitting the need for further ultrasound examination [18]. Indeed, the diagnostic value of D-Dimer testing lies in ruling out DVT because of its highly negative predictive value [18–20]. In conjunction with no high clinical pre-test probability e.g. HemosIL D-Dimer HS 500 assay is accurate when used for DVT diagnostic work-up in out-patients [19, 20].

It is important that selected quality control samples cover a measurement range that include concentrations critical for patient management [21]. The concept of the evaluation of systematic deviation from the target value using the bivariate z-score analysis becomes even more interesting when it distinguishes between results with and without impact on patient management. In the case of the evaluation of D-Dimer this implies the selection of samples around the cut-off level for the exclusion of DVT and samples that have a sufficiently lower or higher D-Dimer concentration than the range around the cut-off level. Therefore, it is also important in an EQA programme to use quality control samples that mimic as most as possible real clinical samples [22–24].

The aim of this study is to develop a practical method for bivariate z-score analysis which can be applied to the survey of an external quality assessment programme. The external quality assessment programme for D-Dimer of the ECAT Foundation has been used to illustrate the application of the proposed methodology.

Materials and methods

EQA surveys

International EQA surveys for D-Dimer are organised by the External quality Control for Assays and Tests (ECAT) organisation (Voorschoten, The Netherlands). In 2022 over 725 laboratories participated in these surveys. The annual programme included four surveys with two control samples in each survey. Control samples were prepared from a single patient donation with an elevated D-Dimer level. After informed consent the patient plasma was collected by plasmapheresis on acid-citrate-dextrose (ACD) [Klinikum Augsburg, Germany]. After centrifugation, samples with different D-Dimer concentrations were prepared by dilution with citrated normal pooled plasma. Plasma samples were lyophilised and all coded samples for a one year's survey programme were distributed by ECAT at room temperature to their participants by postal or courier service. Participants were asked to store the sample until use at 2–8 °C. Under these conditions lyophilised samples are stable for at least 2 years.

Participants were asked to reconstitute each quarter two samples according to a prescribed standard procedure and to measure the D-Dimer level using their standard D-Dimer method. Results were reported to ECAT by a webtool, including information on the method, equipment and unit.

Evaluation of results of the surveys

The evaluation of D-Dimer results includes in total eight samples from the surveys organised in 2022.

After the closing date of a survey all results were evaluated by peer-group analysis. A peer group is defined as a group of participants using the same method for the measurement of D-Dimer. For each peer group with at least 10 participants the robust mean and standard deviation was established using the robust statistical method Algorithm A [6]. Z-scores were calculated on the basis of the robust peer-group means and standard deviation.

Bivariate z-score analysis

Medical laboratories will typically test two EQA control samples at the same time. Consequently, it is expected that the EQA control results will tend to be correlated (rather than independent), a property that will be transferred to their z-score values, within an analytic run, as the correlation coefficient is invariant to the linear transformation that the z-score standardisation procedure performs. In the area of Statistical Process Control & Monitoring (SPC/M) it is well known and documented [13] that when the aim is to monitor the quality of bivariate data, it is more efficient to use a bivariate control chart approach, as opposed to monitoring each dimension separately. The proposed methodology starts by identifying the bivariate outliers, via a Supervised Sequential Hotelling's T^2 control chart. The outlying data are removed, and all the remaining data are used to provide robust estimates of the parameters of the assumed underlying bivariate normal distribution. Based on these estimates two nested homocentric ellipses are drawn, corresponding to confidence levels of 95 and 99.7 %. These ellipses partition the plane into three non-overlapping regions that will indicate the status of the bivariate points as:

- "Acceptable" (green light) if a pair of z-scores lies within the inner ellipse (green region)
- "Need attention" (warning orange light) if a pair of z-scores lie between the two ellipses (orange region) or
- "Unsatisfactory" (red light alarm) if a pair of z-scores lie outside the outer ellipse (red region).

The sample estimate of the overall z-scores mean will tend to be zero and since the EQA control samples are measured at the same time it is expected to have a positive correlation, resulting in ellipses that will tend towards the first diagonal of the axes. The stronger the correlation (i.e., the bigger its absolute value) the narrower these ellipses will be, while on the other extreme of correlation close to zero (indicating independent normal z-scores), the ellipses will turn into circles (see Figure 1).

In Appendix A all the steps of the proposed bivariate z-score analysis, including the technical details are provided in pseudocode.

The proposed methodology analyses bivariate data, and for the assumed bivariate normal distribution five parameters need to be estimated from the data (mean, standard deviation for each dimension and their correlation). Thus, from a statistical perspective using the suggested methodology is not recommended when the volume of available bivariate data is small, as the standard errors of the estimated parameters will be large. The principle is that the greater the amount of data the more accurate the estimate will be (there is no single sample size value that is universally accepted), and from a practical perspective more than 80 or 100 data points appears to be a safe choice.

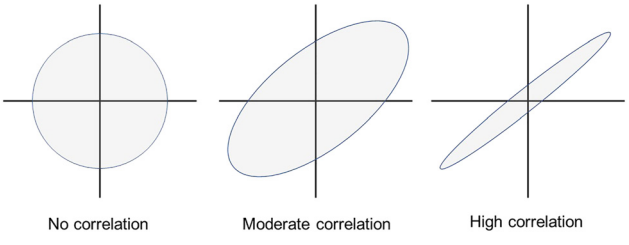

Figure 1: Examples of different levels of correlation in the bivariate z-score plot.

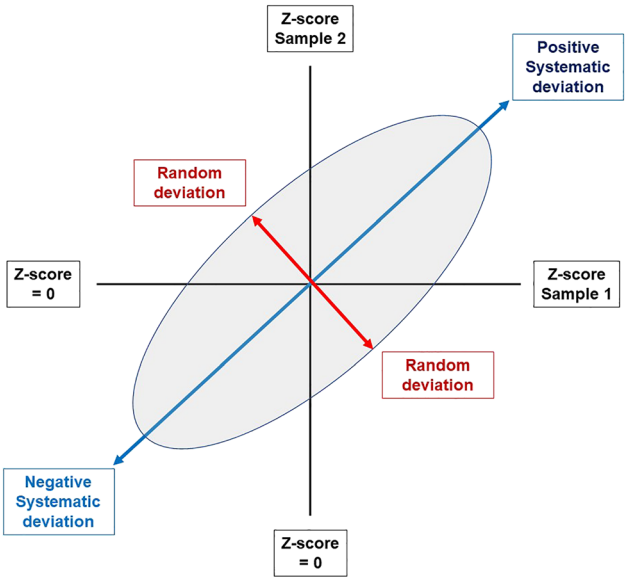

Figure 2: Sample plot of how systematic and random deviation in the bivariate z-score plot can be observed.

The bivariate z-score plot described provides the laboratory with an indication of both systematic and random deviations from zero z-score values [25] (see Figure 2).

Bivariate z-score analysis on the results of surveys

To apply the proposed bivariate z-score analysis the z-scores obtained from the 4 D-Dimer surveys in 2022 were used (survey 2022-D1 – 2022-D4). First for each of the four surveys the overall analysis had to be performed, including the z-score pairs for all methods with at least 10 participants.

In addition, for survey 2022-D4 a separate bivariate z-score analysis was also performed for the three most frequently used methods (A=Siemens Innovance D-Dimer; B=Stago STA-Liatest D-Di Plus; C=Werfen HemosIL D-Dimer HS 500). These three methods cover approximately 80 % of the responses in the ECAT D-Dimer EQA programme (see Table 1).

Results

Survey results

In Table 1 an overview is given of the robust mean and standard deviation of each sample used in the four EQA surveys in 2022 and for each of the peer groups with at least 10 participants. Two samples have been used in two different surveys (sample 2 in survey 2022-D1/sample 1 in survey 2022-D3 and sample 2 in survey 2022-D3/sample 2 in survey 2022-D4).

The proposed bivariate z-score analysis was applied in all four surveys data of Table 1 and the resulting ellipses and data point classification are shown in Figure 3.

Table 1: D-Dimer results (mg/L) for surveys performed in 2022 (mean ± SD).

Survey			2022-D1		2022-D2		2022-D3		2022-D4	
Method	**Unit[a]**	**n[b]**	**Sample 1**	**Sample 2[c]**	**Sample 1**	**Sample 2**	**Sample 1[c]**	**Sample 2[d]**	**Sample 1**	**Sample 2[d]**
IL HemosIL D-Dimer HS	D-Dimer	28	0.30 ± 0.07	0.15 ± 0.03	0.12 ± 0.03	0.32 ± 0.02	0.15 ± 0.02	0.12 ± 0.03	0.19 ± 0.02	0.12 ± 0.03
Radiometer AQT90 Flex D-Dimer	D-Dimer	13	0.65 ± 0.11	0.37 ± 0.04	0.20 ± 0.02	0.67 ± 0.09	0.36 ± 0.03	0.23 ± 0.03	0.28 ± 0.04	0.23 ± 0.03
Sysmex LIAS Auto D-Dimer Neo	D-Dimer	11	1.40 ± 0.13	0.72 ± 0.10	0.77 ± 0.07	1.60 ± 0.12	0.66 ± 0.04	0.49 ± 0.05	1.36 ± 0.12	0.48 ± 0.03
BioMerieux Vidas D-Dimer	FEU	23	0.46 ± 0.11	0.33 ± 0.03	0.24 ± 0.03	0.63 ± 0.05	0.32 ± 0.03	0.24 ± 0.02	0.27 ± 0.02	0.24 ± 0.02
Werfen HemosIL D-Dimer HS 500	FEU	171	1.10 ± 0.15	0.42 ± 0.05	0.37 ± 0.05	1.30 ± 0.10	0.42 ± 0.05	0.31 ± 0.05	0.80 ± 0.14	0.31 ± 0.05
Mindray D-Dimer	FEU	14	1.15 ± 0.11	0.59 ± 0.07	0.50 ± 0.06	1.13 ± 0.14	0.46 ± 0.06	0.43 ± 0.07	0.35 ± 0.04	0.30 ± 0.10
Mitsubishi Pathfast D-Dimer	FEU	10	0.39 ± 0.09	0.42 ± 0.09	0.30 ± 0.02	1.12 ± 0.11	0.40 ± 0.05	0.32 ± 0.05	–	–
Roche Cardiac D-Dimer	FEU	16	0.22 ± 0.05	0.13 ± 0.01	0.10 ± 0.00	0.34 ± 0.04	0.11 ± 0.01	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00
Roche Tinaquant 2nd gen. (cal. citrate)	FEU	17	0.60 ± 0.12	0.30 ± 0.11	0.24 ± 0.12	0.52 ± 0.13	0.31 ± 0.11	0.21 ± 0.12	0.27 ± 0.08	0.20 ± 0.06
Roche Tinaquant 2nd gen. (cal. heparin)	FEU	44	0.54 ± 0.15	0.23 ± 0.06	0.17 ± 0.03	0.44 ± 0.10	0.23 ± 0.06	0.16 ± 0.06	0.22 ± 0.07	0.17 ± 0.06
Siemens Innovance D-Dimer	FEU	381	1.41 ± 0.18	0.70 ± 0.08	0.77 ± 0.10	1.63 ± 0.16	0.64 ± 0.06	0.48 ± 0.05	1.38 ± 0.17	0.46 ± 0.05
Stago Liatest D-Dimer	FEU	22	0.79 ± 0.14	0.42 ± 0.08	0.32 ± 0.03	0.62 ± 0.10	0.40 ± 0.04	0.27 ± 0.03	0.31 ± 0.07	0.28 ± 0.03
Stago STA-Liatest D-Di plus	FEU	108	0.76 ± 0.15	0.43 ± 0.07	0.32 ± 0.06	0.62 ± 0.06	0.40 ± 0.06	0.27 ± 0.04	0.30 ± 0.06	0.27 ± 0.04

[a]D-Dimer methods can express their results in either D-Dimer units or fibrinogen equivalent units (FEU). [b]Numbers are taken from survey 2022-D1. [c]This sample is used in both survey 2022-D1 and 2022-D3. [d]This sample is used in both survey 2022-D3 and 2022-D4.

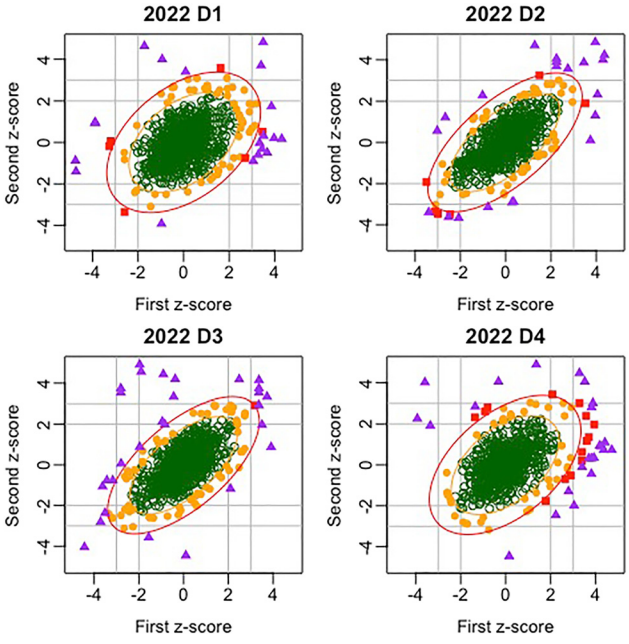

Figure 3: The orange (95 %) and red (99.7 %) ellipses for each of the four 2022 surveys, along with the respective data points. Purple triangles indicate bivariate outliers, red square points refer to alarms, orange-filled discs are the cases that receive a warning and green open circles indicate the conforming cases.

Table 2 shows a series of summary statistics regarding the available data points per survey along with the point estimates of the five parameters of the bivariate normal distribution that were used to construct the respective ellipses.

Table 2: The summary of statistics for the four surveys D1–D4 of 2022.

	2022 – D1	2022 – D2	2022 – D3	2022 – D4
Number of cases	856	837	847	830
Number of univariate outliers	17	14	12	27
Number of bivariate outliers	20	21	28	24
Mean z-score sample 1	0.001	−0.032	−0.006	−0.026
Mean z-score sample 2	−0.005	−0.007	−0.027	−0.011
SD z-score sample 1	0.982	0.991	0.987	0.979
SD z-score sample 2	0.984	0.978	0.979	0.986
Correlation	0.431	0.662	0.647	0.493

Next the suggested bivariate z-score analysis was examined within a survey (2022-D4) across different methods (Figure 4). In more detail, the bivariate analysis was performed for the entire data set and subsequently for each of the subset of data corresponding to the three most frequently used methods (A, B and C; for details see materials and methods). Figure 4 illustrates the fitted ellipses along with the data per case, while in Table 3 the summary statistics of each case are presented.

Discussion

Accurate test results are of major importance in laboratory medicine. This is of specific importance when a measurand

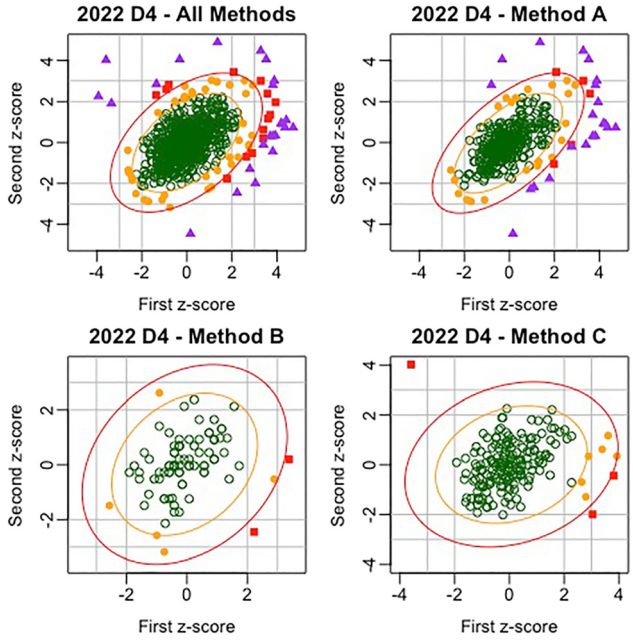

Figure 4: The orange (95 %) and red (99.7 %) ellipses for all the data of survey 4 and each of the individual methods A, B and C, along with the respective data points. Purple triangles indicate bivariate outliers, red square points refer to alarms, orange-filled discs are the cases that receive a warning and green open circles indicate the conforming cases.

Table 3: The summary of statistics for the survey D4 of 2022.

	All methods	Method A	Method B	Method C
Number of cases	830	385	95	184
Number of univariate outliers	27	7	6	4
Number of bivariate outliers	24	21	0	0
Mean z-score sample 1	−0.026	−0.052	−0.075	0.072
Mean z-score sample 2	−0.011	−0.033	0.009	0.008
SD z-score sample 1	0.979	0.980	0.988	1.132
SD z-score sample 2	0.986	0.994	1.058	0.964
Correlation	0.493	0.636	0.278	0.208

is used for medical decision limits [26, 27]. Here we describe this specifically for D-Dimer, a measurand that is used for the exclusion of deep venous thrombosis or pulmonary embolism. Assessment of the (relative) accuracy of measurement is done by participation in an external quality assessment (EQA) programme [5]. The ECAT Foundation provides such a programme for D-Dimer using patient-based control samples. It is important to mimic as closely as possible the clinical laboratory practice. It has been demonstrated that artificially prepared D-Dimer samples behave different from patient-based samples [24].

It can be appreciated from Table 1 that there could be substantial differences between the consensus values of different methods. It is well known that D-Dimer assays used in clinical practice are not standardised nor harmonised [23, 28, 29]. Recently the clinical consequences of these differences have been discussed [19].

To compare the analytical results from a large cohort of laboratories using a variety of different D-Dimer methods in an external quality assessment programme, it is therefore not possible to evaluate the absolute differences between the results of the laboratories and the common or method-specific target value. A harmonised approach to expressing the deviation between the target value and the results of the laboratories should be used. For this purpose, the z-score is used in the programme of the ECAT Foundation. This reflects the deviation corrected by the variation in the distribution of the participants' results [7, 8]. For each method included in the evaluation, the corresponding target value and variation in the distribution (= standard deviation) was established and the corresponding z-score for the results of the laboratories results calculated. This makes comparison between the participants' z-scores of different methods possible.

It has already been demonstrated that in EQA surveys, in which multiple samples has been distributed, only looking in a univariate manner at the z-score performance limited the information that could be extracted from the survey results [9]. This study therefore investigated further how the z-score pairs of all participants could be evaluated in a bivariate manner. It had already been suggested that the correlation of the z-score pairs should be examined [9]. However, it should be realised that correlation assessment is prone to existing outliers [30]. In the current methodology described here we therefore propose starting first with an outlier detection procedure. Initially, a threshold of z-score of –5/5 was used in identifying univariate outliers. This is of course an arbitrary choice, however, the likelihood in a normal distribution of having results outside these thresholds is negligible. The next step is the identification of bivariate outliers by the Supervised Sequential Hotelling's T^2 control chart [14], before the bivariate z-score parameters (means, standard deviations and correlation) have been calculated. This ensures that the assessment position of an individual laboratory in the bivariate z-score analysis is not affected by apparent outliers. The z-score pair distribution after the removal of outliers reflects the state-of-the-art performance of participating laboratories, considering the deviation from the target values for both samples. This is called the bivariate z-score analysis.

After the outlier removal, the distribution of the z-score pairs was assessed, calculating the 95 and 99.7 % confidence borders of the distribution. This is a comparable approach as the one used in a univariate z-score analysis, where the

z-score ranges from −2 to 2 and from −3 to −2 or 2 to 3 are used for performance assessment [6]. The 95 % distribution area is reflected by the green symbols in Figures 3 and 4. The distribution between 95 and 99.7 % is reflected by the orange filled disc symbols in these figures, while the z-score pairs outside the 99.7 % distribution are indicated by red square symbols. Outlying z-score pairs are identified by purple triangle symbols. In the survey reports, the position of a participant is indicated by a black open circle. This gives the participant the advantage of seeing their position within the state-of-the-art distribution. It also helps the laboratory to assess whether any corrective actions are required.

Figure 3 and Table 2 show the distribution of the z-score pairs and the corresponding correlation parameters for the four different D-Dimer surveys in 2022. All four surveys show a moderate correlation (0.431–0.662) with no significant differences between the four surveys independently of the combination of D-Dimer samples used. This indicates a relatively stable state-of-the-art situation in laboratory performance.

In Figure 4 the bivariate z-score analysis for the three most frequently used methods was further investigated. For this purpose, the results of the survey 2023-D4 were selected because from two frequently used methods (Werfen HemosIL D-Dimer HS 500 [method C] and Siemens Innovance D-Dimer [method A]), this survey includes a sample with an elevated D-Dimer level (sample 1) and one with a D-Dimer level slightly below the medical decision limit of 0.5 mg/L. This is not seen for the third frequently used method (Stago STA-Liatest D-Di Plus [method B]), which indicates the potential heterogeneity in response between different D-Dimer methods. However, the bivariate z-score correlation is not affected by the differences in D-Dimer levels measured between the methods (Table 3). For users of the Werfen and Siemens methods, approx. 65 % of all participants in survey 2022-D4, thus can assess how well their measurement system distinguishes patients with a high risk of thromboembolism vs. those for whom thromboembolic events can be ruled out. In contrast, for methods with D-Dimer levels below medical decision limit in both samples, such an evaluation cannot be made. Ideally samples should be used with similar features in all methods. However, because of the heterogeneity in the responsiveness between D-Dimer methods and the available source of control samples this will be practically difficult to achieve.

When all methods are combined into a single data set then the sample size increases significantly, leading to estimates with smaller uncertainty, which are therefore more trustworthy. The only underlying assumption in merging all methods together is that the z-score-generating mechanism is the same across different methods. On the basis of the statistical results presented in Table 3 there is no reason to assume there is a difference in the z-score distribution between methods.

The strength of the bivariate z-score analysis relies on the fact that it takes into account the correlation between the z-score pairs, something that is ignored in a univariate evaluation. In the latter we derive z-score rectangular acceptance limits (based on the z-score limits [−2, 2] and [−3, 3]) as opposed to the actual elliptically shaped borders [13]. Notably, even when the two z-scores are independent (i.e. their correlation is zero) the proper acceptance limits are circular and not rectangular. A weakness of the bivariate z-score analysis is the fact that it is more complicated.

Important lessons can be drawn from the bivariate z-score analysis by the laboratories participating in an EQA programme. First, the bivariate z-score analysis expresses the state-of-the-art situation with respect to systematic and random deviations from the assigned values. The narrower the ellipse, the more systematic deviations drives the position of a laboratory within the distribution of z-score pairs. The broader the ellipse, the more random deviations may play a role (see Figure 2). Secondly, the bivariate z-score analysis indicates whether the z-score pair of a laboratory is positioned within or outside the 95 % confidence limits. If a laboratory is positioned outside the 95 % confidence limits it is a warning signal that corrective action might be required. This is even more pronounced when a laboratory is positioned outside the 99.7 % confidence limits (red zone). A third lesson which can be drawn focuses on the position of the z-score pair belonging to a laboratory. If the laboratory is positioned in the left lower quadrant (see Figure 2) it means there is a negative systematic deviation from the assigned value. On the other hand, if a laboratory is positioned in the upper right quadrant there is a positive systematic deviation from the assigned value. If a laboratory is positioned in the upper left or lower right quadrants it means that the z-score does not have the same sign. If the z-scores are low, this may indicate that the position could be driven by acceptable random deviations. However, when the z-scores are high this could indicate unacceptable random errors and corrective actions are needed, even if the position is still within the 95 % confidence limits. It should be realized that opposite signs for the z-scores could also be caused by proportional bias. It is therefore always the responsibility of a laboratory to carefully investigate the potential cause for deviating results [31].

The magnitude of acceptable random deviation depends on the measurand, the level of the measurand and whether this level is close to clinical decision limits. In the example described here, it is important to evaluate whether there is an increased risk of accidently reporting a false negative or positive D-Dimer result. This may have significant clinical

implications, and should also be considered when a z-score pair is positioned in the lower left or upper right quadrant (Figure 2). If the negative or positive systematic deviation is too great, in this case also a false negative, or positive D-Dimer results could be reported. If this is so, there is a systematic deviation, which implies that all the samples measured have a certain systematic deviation. The bivariate z-score analysis may advantage the laboratory community by highlighting potential systematic and random deviations and supporting the required corrective actions. Finally, this may lead to optimally reliable test results which are of major importance for patient care. The importance to use performance criteria that ensure that laboratory results are produced with sufficient quality for patient care have been emphasized [32]. One approach is the use of performance limits defined by external quality assessment specifications. Here we described an approach based on modern statistical process control principles applied to external quality control results. This may advance the clinical laboratory community given the fact that clinical relevant samples are used in an external quality assessment programme.

In conclusion, implementation of the bivariate z-score analysis is of major importance when an EQA provider distributes around multiple samples in the same survey and dependency between the reported results is plausible. Important lessons can be drawn from the shape of the ellipse with respect to random and systematic deviations. In addition, individual laboratories are informed about their position in the state-of-the-art distribution and whether they need to deal with systematic and/or random deviations.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission. All authors contributed equally to this work.
Competing interests: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

Appendix A: Algorithm to implement the bivariate z-score analysis

1. **Read the data set**: Each data point is a pair of z-scores (2×1 vector) from a given survey. We will denote the data point (i) by:

$$\boldsymbol{z}_i = (z_{1i}, z_{2i})$$

where, z_{1i}, z_{2i} refer to the first and second z-score of case (i) in the data set of the specific survey.

2. **Clean the data from missing values**: If for a pair of z-scores ($\boldsymbol{z}_i$), at least one of the two z-score values is a missing value (NA) then we remove this pair from the data set.
3. **Clean the data from univariate outlying z-scores**:We set a positive threshold value Z* (for example Z*=5) and if the absolute value of any z-score individual value (i.e., z_{1i} or z_{2i}) exceeds Z*, then the pair that contains the outlying observation is removed.
4. **Clean the data from bivariate outlying z-scores**: we will adopt a Supervised Sequential Hotelling's T^2 statistic control chart to identify candidate bivariate outliers:

   Functions needed in this step:

   – **Sample mean vector**: a 2×1 vector estimated by:

$$\overline{\boldsymbol{z}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i$$

where, $\boldsymbol{z}_i$ for i=1,2, …, n are the n available data points

   – **Sample covariance matrix**: a 2×2 matrix estimated by:

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_i - \overline{\boldsymbol{z}})'$$

where, $\boldsymbol{z}_i$ for i=1, 2, …, n are the n available data points (the prime symbol denotes the transposing of the vector). $\boldsymbol{S}$ will be a symmetric matrix and we will denote it from now on as:

$$S = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

   – **Hotelling's T^2 statistic**: estimated for each data point z by:

$$T^2 = (\boldsymbol{z} - \overline{\boldsymbol{z}})' \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{z}})$$

where the prime denotes the transposed vector and $\boldsymbol{S}^{-1}$ is the inverse of the matrix $\boldsymbol{S}$. So for each of i=1,2, …, n we will derive T_i^2, i.e. we will have one score for each data point.

   – **Hotelling's Upper Control Limit (UCL)**: the control chart's limit which will be used to identify candidate bivariate outliers. It is estimated by:

$$UCL = \frac{(n-1)^2}{n} Beta\left(\alpha; 1, round\left(\frac{n-3}{2}\right)\right)$$

where $Beta\left(\alpha; 1, round\left(\frac{n-3}{2}\right)\right)$ is the upper α percentage point of a beta distribution with parameters 1 and $round\left(\frac{n-3}{2}\right)$. For the value of α it is suggested to use $\alpha = 0.0027$.

End of functions definition

We start with the n data-point pairs, and we will identify and remove sequentially bivariate outliers using a loop that will check two conditions for each iteration:

Condition 1: for at least one data point we have $T_i^2 > UCL$

Condition 2: The minimum of the diagonal elements of matrix $\boldsymbol{S_i}$ (i.e. the estimated sample covariance matrix, where the subscript i indicates the value i that was left out from the estimation and for which condition 1 is satisfied) exceeds K, i.e. $\min(a,c)>K$, where K is a number near but smaller than 1 (e.g. K=0.95). Since we model z-scores the variance should be 1 and so this condition helps to certify that after a candidate outlier is removed the variance does not become too small compared to the target value of 1.

The steps for removing the bivariate outliers are as follows:

I. Start with the available data points (at the end of step **3** where univariate outliers were cleaned).
II. Estimate from the available data $\overline{\boldsymbol{z}}$ (sample mean vector)
III. Estimate from the available data $\boldsymbol{S}$ (sample covariance matrix)
III. Estimate for each of the available data point (i) the T_i^2 (Hotelling's T^2 statistic score)
IV. Estimate the current UCL (Hotelling's T^2 Upper Control Limit)
V. When (**Condition 1 & Condition 2** are **TRUE**) do the loop:
   a. Identify the $\max_i(T_i^2)$. The data point $\boldsymbol{z}_i$ is judged as a bivariate outlier and is removed from the data set.
   b. On the reduced data set of the last step do the estimation steps II, III, IV, V.

End the loop.

5. Assuming bivariate normality, we make use of the estimated mean ($\overline{\boldsymbol{z}}$) and variance-covariance matrix ($\boldsymbol{S}$) from the last iteration of the previous step (i.e., after the univariate and bivariate outliers are removed) and draw two homocentric ellipses corresponding to confidence of 95 % (orange colour, using as critical value $\chi^2_{2;0.05} = 5.991$) and 99.7 % (red colour, using as critical value $\chi^2_{2;0.0027} = 11.829$), where the orange will be nested in the red ellipse.
6. On the plot of the two homocentric ellipses (orange and red) we superimpose the data points and we use a distinct symbol to denote the outlying observations. Depending on where a point falls, we characterise it as conforming or non-conforming. In the latter case we raise the alarm, which is either orange or red. Precisely, for each data point one of the following is true:
   – **No alarm**: if the point is plotted within the orange ellipse.
   – **Orange alarm**: if the point is plotted outside the orange ellipse but inside the red ellipse (i.e. it does not conform to the 95 % region, but it conforms to the 99.7 % region).
   – **Red alarm**: if the point is plotted outside the red ellipse (i.e. it does not conform to the 99.7 % region).

References

1. Rohr UP, Binder C, Dieterle T, Giusti F, Messina CG, Toerien E, et al. The value of in vitro diagnostic testing in medical practice: a status report. PLoS One 2016;11:e0149856.
2. Hicks AJ, Carwardine ZL, Hallworth MJ, Kilpatrick ES. Using clinical guidelines to assess the potential value of laboratory medicine in clinical decision-making. Biochem Med 2021;31:010703.
3. Epner PL. Appraising laboratory quality and value: what's missing? Clin Biochem 2017;50:622–4.
4. Sandars J, Esmail A. The frequency and nature of medical error in primary care: understanding the diversity across studies. Fam Pract 2003;20:231–6.
5. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232–5.
6. ISO:13528. Statistical methods for use in proficiency testing by interlaboratory comparison. Geneva: International Organization for Standardization (ISO); 2022.
7. Sareen R. Illuminating Z-score in external quality assessment for medical laboratory. Health Care Curr Rev 2018;6:1000228.
8. Coucke W, Soumali MR. Demystifying EQA statistics and reports. Biochem Med 2017;27:37–48.
9. Coucke W, Rida Soumali M, Badrick T. Improving Youden plots by including analytical performance specifications. Clin Chim Acta Int J Clin Chem 2022;531:212–6.
10. Tracy ND, Young JC, Mason RL. A bivariate control chart for paired measurements. J Qual Technol 1995;27:370–6.
11. Shirono K, Iwase K, Okazaki H, Yamazawa M, Shikakume K, Fukumoto N, et al. A study on the utilization of the Youden plot to evaluate proficiency test results. Accred Qual Assur 2013;18:161–74.
12. Zhou Q, Hu J, Li X, Li S, Gao Z, Xie W, et al. Comparison of traditional, trimmed traditional and robust Youden charts. Clin Chim Acta Int J Clin Chem 2015;446:213–7.
13. Montgomery DC. Introduction to statistical quality control, 8th ed. New York: John Wiley & Sons, Inc.; 2020:458–87 pp.
14. Hotelling H. Multivariate quality control illustrated by air testing of sample bombsights. In: Eisenhart C, Hastay MW, Wallis WA, editors. Techniques of Statistical Analysis. New York: McGraw-Hill; 1947:111–84 pp.
15. Mason RL, Young JC. Multivariate statistical process control with industrial applications. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM); 2002.
16. Dechert J, Case KE. Multivariate approach to quality control in clinical chemistry. Clin Chem 1998;44:1959–63.
17. Johnson ED, Schell JC, Rodgers GM. The D-dimer assay. Am J Hematol 2019;94:833–9.
18. Oude Elferink RF, Loot AE, Van De Klashorst CG, Hulsebos-Huygen M, Piersma-Wichers M, Oudega R. Clinical evaluation of eight different D-dimer tests for the exclusion of deep venous thrombosis in primary care patients. Scand J Clin Lab Invest 2015;75:230–8.

19. Hamer HM, Stroobants AK, Bavalia R, Ponjee GAE, Klok FA, van der Hulle T, et al. Diagnostic accuracy of four different D-dimer assays: a post-hoc analysis of the YEARS study. Thromb Res 2021;201: 18–22.
20. Legnani C, Cini M, Scarvelis D, Toulon P, Wu JR, Palareti G. Multicenter evaluation of a new quantitative highly sensitive D-dimer assay, the Hemosil D-dimer HS 500, in patients with clinically suspected venous thromboembolism. Thromb Res 2010;125:398–401.
21. Sobas F, Mazliak L, Bellisario A, Lefranc M, Lienhart A, Nougier C, et al. Determining the adequate number of internal quality control levels: the example of coagulation factor VIII assay. Blood Coagul Fibrinolysis 2008;19:433–7.
22. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52:553–4.
23. Meijer P, Haverkate F, Kluft C, de Moerloose P, Verbruggen B, Spannagl M. A model for the harmonisation of test results of different quantitative D-dimer methods. Thromb Haemostasis 2006;95: 567–72.
24. Bevan S, Longstaff C. Is it possible to make a common reference standard for D-dimer measurements? Communication from the ISTH SSC Subcommittee on Fibrinolysis. J Thromb Haemostasis 2022;20: 498–507.
25. Martín J, Velázquez N, Asuero AG. Youden two-sample method. In: Kounis LD, editor. Quality control and assurance – an ancient Greek term re-mastered. InTech; 2017:47–83 pp.
26. Werner M. Linking analytic performance goals to medical outcome. Clin Chim Acta Int J Clin Chem 1997;260:99–115.
27. Horvath AR, Bossuyt PM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WD, et al. Setting analytical performance specifications based on outcome studies – is it possible? Clin Chem Lab Med 2015;53: 841–8.
28. Longstaff C, Adcock D, Olson JD, Jennings I, Kitchen S, Mutch N, et al. Harmonisation of D-dimer – a call for action. Thromb Res 2016;137: 219–20.
29. Thachil J, Longstaff C, Favaloro EJ, Lippi G, Urano T, Kim PY, et al. The need for accurate D-dimer reporting in COVID-19: communication from the ISTH SSC on fibrinolysis. J Thromb Haemostasis 2020;18:2408–11.
30. Ellison SLR. Applications of robust estimators of covariance in examination of inter-laboratory study data. Anal Methods 2019;11: 2639–49.
31. Kristensen GB, Meijer P. Interpretation of EQA results and EQA-based trouble shooting. Biochem Med 2017;27:49–62.
32. Klee GG. Establishment of outcome-related analytic performance goals. Clin Chem 2010;56:714–22.