Available online at www.sciencedirect.com

SCIENCE DIRECT®

Clinica Chimica Acta 368 (2006) 160 – 167

www.elsevier.com/locate/clinchim

Trueness verification and traceability assessment of results from commercial systems for measurement of six enzyme activities in serum An international study in the EC4 framework of the Calibration 2000 project

Rob Jansen [a,*], Gerhard Schumann [b], Henk Baadenhuijsen [c], Paul Franck [d], Carlo Franzini [e], Rolf Kruse [f], Aldy Kuypers [c], Cas Weykamp [g], Mauro Panteghini [e]

[a] *Department of Clinical chemistry, St. Anna Hospital, Post box 90, 5660 AB Geldrop, The Netherlands*
[b] *Department of Clinical Chemistry, Medizinische Hochschule Hannover, Hannover, Germany*
[c] *Dutch Foundation for Quality Assessment in Clinical Laboratories, Nijmegen, The Netherlands*
[d] *Haga Hospital, The Hague, The Netherlands*
[e] *Department of Clinical Sciences "Luigi Sacco", University of Milan, Milan, Italy*
[f] *DGKL, Referenzinstitut für Bioanalytik, Bonn, Germany*
[g] *Queen Beatrix Hospital, Winterswijk, The Netherlands*

Received 21 December 2005; accepted 22 December 2005
Available online 6 March 2006

Abstract

Background: The in vitro diagnostics directive of the European Union requires traceability to higher order reference measurement procedures and materials for analytes in assuring the result trueness and comparability of laboratory measurements. Manufacturers must ensure that the systems they market are calibrated against available reference systems. Validation of metrologically traceable calibrations is, however, required.

Methods: A commutable serum-based material was analyzed in three reference laboratories and target values for six enzymes (ALT, AST, CK, GGT, LD, amylase) were assigned using IFCC reference measurement procedures. 70 laboratories in Germany, Italy, and The Netherlands measured the same enzymes in the material using procedures from six commercial companies. A system for maximum allowable error was developed from the biological variation model and the results of the various procedures were tested on their compliance to trueness and between-laboratory and within-laboratory variations relative to the maximum allowable.

Results: For ALT results were relatively good. >95% of laboratories using systems from Dade, Olympus, Ortho and Roche are expected to comply traceability within the biologically derived limits, and 94% respectively 89% from Abbott and Beckman. For AST and GGT only Dade respectively Olympus fully complied. For CK all companies showed significant bias. Nevertheless >95% of laboratories applying Abbott, Beckman and Roche systems will comply. Finally, LD and amylase measurements require significant improvement. Some manufacturers continue to sell on the European market assays giving results which are not traceable to the internationally accepted reference systems.

Abbreviations: EC4, European Communities Confederation of Clinical Chemistry and Laboratory Medicine; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; IVD, in vitro diagnostic; ALT, alanine aminotransferase; AST, aspartate aminotransferase; CK, creatine kinase; GGT, γ-glutamyl transferase; LD, lactate dehydrogenase; EQAS, external quality assessment scheme; T, target value; SD_{bl}, between-laboratory variance; SD_{wl}, within-laboratory variance; SD_{bl} max, maximum allowable between-laboratory variance; SD_{wl} max, maximum allowable within-laboratory variance; AB, desirable analytical bias; SD_d, desirable analytical variation; SD_w, within-person biological variation; SD_b, between-person biological variation; TAE, total allowable error.

* Corresponding author.
E-mail address: r.jansen@st-anna.nl (R. Jansen).

0009-8981/$ - see front matter
doi:10.1016/j.cca.2005.12.033

Conclusions: The traceability of enzyme measurements obtained with routine procedures to internationally accepted IFCC reference systems is not yet satisfactorily accomplished in clinical practice.

Keywords: Enzymes; Reference methods; Calibration; Bias; Analytical error; IVD; Verification; External quality assessment; Routine methods

1. Introduction

The European directive 98/79 IEC on in vitro diagnostic (IVD) medical devices requires the traceability of results obtained in clinical laboratories to internationally recognized and accepted reference materials and reference measurement procedures in assuring accuracy and comparability of measurements [1,2]. In order to achieve metrological traceability, a hierarchical system of analytical procedures with decreasing measurement uncertainty and increasing trueness has to be established [3].

The catalytic concentration of enzymes build a special class of quantities, defined in terms of the amount of an agreed-upon substrate they convert in an agreed-upon measurement system, the so-called "catalytic amount". The way to achieve standardization of routine measurements for these types of quantities is described in the ISO 18153 standard via the development of International Conventional Reference Measurement Procedures, here referred to as IFCC primary reference measurement procedures [4]. In fact, oppositely to other measurands (or quantities), the numerical results of catalytic activity measurements depend entirely on the experimental conditions under which the measurements are made. In the standardization of enzyme assays, therefore, a reference measurement procedure, which defines the conditions under which a given enzyme activity is measured, occupies the uppermost position filled, with regard to non-enzyme analytes, by primary reference materials [5]. IFCC primary reference measurement procedures are currently available for alanine aminotransferase (ALT, EC 2.6.1.2); aspartate aminotransferase (AST, EC 2.6.1.1); creatine kinase (CK, EC 2.7.3.2); γ-glutamyltransferase (GGT, EC 2.3.2.2); lactate dehydrogenase (LD, EC 1.1.1.27) [6–10]; and α-amylase (EC 3.2.1.1) (Schumann G, Aoki R, Ferrero CA, Ehlers G, Férard G, Gella FJ, Jørgensen PJ, Kanno T, Kessner A, Klauke R, Kytzia HJ, Lessinger JM, Nagel R, Pauwels J, Schimmel H, Siekmann L, Weidemann G, Yoshida K, Ceriotti F. IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C. International Federation of Clinical Chemistry and Laboratory Medicine. Part 8. Reference procedure for the measurement of catalytic concentration of alpha-amylase, *manuscript in preparation*).

Adequate calibration of commercial systems, through well validated and certified calibrators, may assure values which are all traceable to these IFCC primary reference measurement procedures, then giving comparability of patient data, even when the same patient is examined in healthcare institutions using different commercial systems [5].

Calibration in serum enzyme measurements has long been recognised as a means to reduce the existing between-laboratories variances [11]. However, calibration by individual laboratories using secondary certified reference materials would be very costly and practically not feasible because of the large batches of materials needed. The correct implementation of the reference system concept required in the pertinent ISO standard should solve this problem [4,5]. However, it is the responsibility of the profession to verify the IVD trueness and compatibility of the methods by performing trueness verification studies of the instruments and methods used in daily practice [12,13]. This should be a task for the external quality assessment scheme (EQAS) organisers through the use of commutable materials with values assigned by IFCC primary reference measurement procedures.

The Calibration 2000 project of the Netherlands aims at the development of commutable certified materials, i.e. materials having values assigned by reference measurement procedures that are proven to behave as patient samples, which can be used to assign traceable values to commercial calibrators [14–18]. Clinical laboratories, which will use routine procedures with these validated calibrators to measure human patient specimens, may finally obtain values which are traceable to the reference system. Then, the traceability requirement, as formulated by the IVD directive of the European Union, can be implemented in practice.

The commutable materials developed in the Calibration 2000 project can also be useful for the trueness verification, i.e. the assessment of the entity of bias of results obtained by a routine method when compared with the true value for a given analyte. In the project it was demonstrated that the lack of commutability in the materials usually employed in EQAS did not permit this type of estimate [18]. Materials were then developed for lipids and enzymes that were proven to be commutable and useful for trueness verification [16–18].

The present paper reports the results from a pilot study performed in three European countries (Germany, Italy, and The Netherlands) to assess trueness and comparability of results of the most frequently used instruments for measurements of serum enzymes for which IFCC primary reference measurement procedures are currently available. The commutable Calibration 2000 material was targeted by reference laboratories using these procedures and then assayed by a total of 70 European laboratories in the three countries. Results from commercial methods were assessed using a system for maximum allowable error derived from

the desirable analytical performance based on the biological variation model [19,20], including a set of three criteria: a) trueness, i.e. bias relative to the target value; b) allowable between-laboratory variation; and c) allowable within-laboratory variation.

2. Materials and methods

2.1. Characteristics of the trueness verification material

The trueness verification material was prepared from fresh human serum pools enriched with recombinant human enzymes, aliquoted in 1 mL vials, and stored frozen at −80 °C. Enzymes used for enrichment of the baseline serum pools were obtained from Asahi Chemical Industry Co., Ltd., Tokyo, Japan. ALT, AST and CK originated from recombinant E. Coli transfected with human liver and muscle genes, GGT from recombinant mammalian cells with human liver gene, and LD from human red blood cells. Amylase was present natively and not spiked. According to the manufacturer's statement, all enzyme preparations have an isoenzyme composition similar to those of the naturally occurring enzymes in human serum. In the Calibration 2000 project materials were tested for commutability using the so-called Twin-Study design [16]. Briefly, 30–40 pairs of laboratories using various methods and instruments were asked to exchange 10 fresh patient samples selected by each pair member with enzyme activities covering the clinically relevant range, and to analyse the 20 patient samples in one run together with the candidate trueness verification material for serum enzymes. Regression lines were constructed through the results for the patient samples for each pair of laboratories, and the distance to the line of the results for the candidate trueness verification material was calculated. If the distance averaged as less than 3 state-of-the-art within-laboratory SD, the material was considered commutable with patient samples across the corresponding routine methods.

2.2. Study protocol

Vials of the selected commutable trueness verification material were sent on dry ice by courier to the reference laboratories for value targeting and to the national EQAS co-ordinators for distribution to the participating clinical laboratories in each country. The material was first analysed in three reference laboratories from the IFCC enzyme network using IFCC primary reference measurement procedures, by a protocol involving, for each enzyme, two determinations per day for three different days. The mean of the results obtained by the reference laboratories was assigned as the target value (T).

Through the national EQAS co-ordinators of the three participating countries, the trueness verification material was then distributed to a total of 70 laboratories, 30 in Germany, 26 in Italy, and 14 in The Netherlands, applying various commercial analytical procedures. Companies (and instruments) included Abbott (Aeroset $n=5$, Architect C8000 $n=4$), Beckman (Synchron LX20 $n=9$, CX9 $n=2$), Dade Behring (Dimension RxL $n=11$), Olympus (AU 400 $n=1$, 640 $n=5$, 2700 $n=3$, 5400 $n=1$), Ortho Clinical Diagnostics (Vitros 250 $n=5$, 950 $n=6$), and Roche (Hitachi 917 $n=4$, Integra 800 $n=3$, Modular $n=11$). The participating laboratories were instructed to measure the material exactly according to the company instructions (e.g. by excluding any correction factors) and were asked to provide information on reagents, wavelength, calibrator, and any issue possibly influencing the measurement. The material was analysed in each laboratory in five replicate measurements in one run for each enzyme. Results were then sent to the organisers and the statistical evaluation centralized.

2.3. Statistical methods

The five intra-laboratory measurements for each enzyme were first inspected for outliers. Next, for each laboratory the average value for each enzyme and the corresponding SD were calculated. These averages formed the basis for further calculations. The laboratory mean values were grouped per company. The pooled within-laboratory SDs were also calculated from the measurements per laboratory obtained for a company for each enzyme. Since the trueness verification material was targeted using IFCC primary reference measurement procedures, plots against T were constructed for each enzyme for each company. Analytical performance was assessed at three levels: bias relative to T, between-laboratory variance (SD_{bl}), and within-laboratory variance (SD_{wl}). The SD_{bl} was tested against the maximum allowable between-laboratory variance (SD_{bl} max) and the pooled SD_{wl} against the maximum allowable within-laboratory variance (SD_{wl} max).

SD_{bl} max and SD_{wl} max were computed from the biological variance-based model of Fraser [19]. In the calculation the desirable analytical variation parameters for individual laboratories were used, as published by Ricos et al. [20]. In this model, for each individual laboratory, the desirable analytical bias (AB) is $<0.25\ [(SD_w)^2+(SD_b)^2]^{0.5}$ and the desirable analytical variation (SD_d) is $<0.5\ SD_w$, where SD_w is the within-person biological variation and SD_b is the between-person biological variation. The total allowable error (TAE) for a single laboratory is $1.65\ SD_d+AB$.

SD_{bl} max is defined as the between-laboratory variation at which the laboratory mean values of 95% of the laboratories are within the AB limit. In the presence of significant bias of the method mean, SD_{bl} max is computed from a one sided probability, in absence of bias from a two-sided. If 95% (one sided approach) of the laboratories applying a method should fulfil these criteria, then $(1.65\ SD_{bl}+|X-T|)<AB$, where SD_{bl} is the between-laboratory

variation for the method, X is the method mean, and T is the target value. It follows that the SD_{bl} max is $(AB-|X-T|)/1.65$ if X is deviating significantly. SD_{bl} max is AB/1.96 (two-sided) if X is not significantly deviating from T. The deviation of X from T was tested using the t-statistic, considering T as a norm.

The SD_{wl} max was computed from $(TAE-1.96\ SD_{bl}-|X-T|)/1.96$ if X is deviating significantly from T, and $(TAE-1.96\ SD_{bl})/1.96$ if X is not significantly deviating from T. The squares of SD_{bl} and SD_{wl} were statistically tested against the squares of SD_{bl} max and SD_{wl} max using the F-test.

The percentage of laboratories that is expected to measure within the allowable area was calculated by estimating the parameter a, as $[(T+1.96\ SD_{bl}\ max)-X]/SD_{bl}$, and the parameter b, as $[(X-(T-1.96\ SD_{bl}\ max)]/SD_{bl}$. In the table of the normal distribution, the densities $D(a)$ at a and $D(b)$ at b can be found. The subtraction $|D(a)-D(b)|\times 100$ gives the percentage of laboratories expected to measure within the established limits. The maximum allowable probability to measure outside the criteria using a given analytical procedure was set to 0.05, and thus to fulfil the criteria this percentage should be >95%.

3. Results and discussion

The Calibration 2000 material used in this study is a deep frozen (−80 °C) pooled serum, which has been spiked with human recombinant enzymes. This material (Dutch enzyme calibrator) was previously proven to be commutable for the commercial methods measuring enzymes considered in the present study [18]. The material was first analysed by three reference laboratories of the IFCC enzyme network. The means (±S.D.) of results for each enzyme obtained in these laboratories were 108.3 U/L (±2.5) for ALT, 184.0 U/L (±1.7) for AST, 302.7 U/L (±1.5) for CK, 155.3 U/L (±1.5) for GGT, 351.3 U/L (±2.5) for LD, and 70.5 U/L (±2.1) for α-amylase, respectively. After the collection of all results from the 70 participating laboratories, the five replicates per laboratory per enzyme were first inspected for outliers. No gross errors were present in the sets of results per enzyme per laboratory. Later on, information and findings received from each laboratory were reviewed and compared with each other to underline possible aspects significantly influencing study results. Data for CK and GGT showed no outlier results or apparent abnormal issues in the reported method settings. However, for GGT the results of one laboratory using Beckman analyzer (mean value, 123.4 U/L) appeared to be markedly different from the method-mean. Although no clear reason was found for this deviation, this result was excluded. Consequently, all data ($n=70$) for CK and 69 results for GGT were used for calculations. For ALT and AST measurements two laboratories used analytical procedures without pyridoxal-5-phosphate and the results were excluded from the calculations. The results of four laboratories that measured pancreatic amylase instead of total amylase were also excluded from the calculations.

LD showed some peculiar issues that needed clarification before to start calculations and statistical evaluations. First of all, the frequency distribution of LD results was clearly bimodal, with 61 laboratory results distributed around a median value of 345 U/L and 9 laboratory results distributed around a median of 850 U/L. In particular, results of all Ortho Clinical Diagnostics systems working in Italy and The Netherlands (7 laboratories in total) were approximately 2.4-times higher than the established T, as a result of the use of a method displaying different analytical specificity for LD [21] and, consequently, determining results that are not traceable to the internationally accepted reference system. In Germany ($n=4$), Ortho apparently used a completely different analytical approach, with results closer to the target. The results of the Dutch and Italian laboratories using Ortho systems were therefore excluded from the following calculations. It should, however, be pointed out that this "country-based" different approach by Ortho appears to be unacceptable from the metrological point of view. To fulfil IVD directive of the European Union, companies should currently refer their results to the higher order, internationally accepted IFCC primary reference measurement procedure [10]. From the data collected in the present study, Ortho is possibly not fulfilling IVD directive at this time in Italy and in The Netherlands. Two other outlier laboratories were identified from LD results. One laboratory using a Beckman procedure produced results around 780 U/L, the reason for which could, however, not be found. Another laboratory using a Roche system gave a mean result of 634 U/L. Though in conflict with the European IVD requirements, Roche currently attributes different values to its internal calibrator for LD assay. In addition to values fulfilling traceability towards the higher order reference measurement procedure, the company still keep calibrator values referring to the pyruvate to lactate assay principle, permitting laboratories to choose between the two values when instruments are calibrated. The results of these two additional laboratories were removed as well and excluded from the final calculations. After the exclusion of 9 outlier results, a unimodal frequency distribution for LD results was obtained.

After these preliminary examination and exclusions, overall means per enzyme per company were then calculated using means for each laboratory employing reagents and instruments from the same manufacturer. The overall means and pooled between- and within-laboratories SDs for each company were tested against T, SD_{bl} max and SD_{wl} max, respectively (see Section 2). Since the laboratories were asked to measure the material in five replicate measurements in one run, the pooled within-laboratory SD represents the within-run SD, thus excluding within-laboratory between lot and between calibration variation.

As the laboratories applied their analytical systems exactly according to the manufacturers' instructions, the between-laboratory variation was expected to be influenced predominantly by between lot and between calibration variance.

Table 1 summarizes the obtained results. For ALT the desirable analytical performance parameters for individual laboratories at T of 108.3 U/L were ±13.0 U/L for AB, <13.2 U/L for SD_d, and ±34.8 U/L for TAE. The mean values per company across countries differed significantly from T for Beckman and Ortho systems. Since there was significant bias for these companies, the SD_{bl} max was smaller than for the companies showing no bias. Although for Beckman and Roche systems, the SD_{bl} was larger than the SD_{bl} max, this was not statistically significant. The between- and within-laboratory variations for other companies also fulfilled the analytical performance criteria for ALT. In particular, the SD_{wl} was significantly smaller than the SD_{wl} max for all of the evaluated companies.

Fig. 1 depicts the target value, the mean values and the SD_{bl} for each company system measuring ALT. In addition, the area of SD_{bl} max when no bias is present is indicated. As can be seen, the distribution of results for Abbott and Roche systems are partly outside the area of maximum allowable SD_{bl}. For Beckman systems the SD_{bl} and the significant bias added to violation of the criteria. Dade and Olympus systems did not show significant bias and the SD_{bl} and SD_{wl} were small enough to fulfil the criteria.

Based on the results of this study, it is to be expected that >95% of laboratories using the methods of Dade, Olympus, Ortho and Roche measure ALT in conformity with the desirable analytical performance based on the biological variation model. 94% of the laboratories using analytical procedures of Abbott, 89% of Beckman users can be expected to measure within these limits.

For AST, the desirable performance parameters for individual laboratories at T of 184.0 U/L were ±9.9 U/L for AB, <11.0 U/L for SD_d, and ±28.0 U/L for TAE, respectively. The overall means per company showed significant bias from T for Dade, Olympus, and Roche systems (Table 1). However, the statistical significance of bias was greatly influenced by the varying size of SD_{bl}. Particularly, the high SD_{bl} of Beckman methods derived from the results of two Dutch laboratories, which had mean AST values of 153.4 U/L and 212.6 U/L, both differing widely, though not statistically outlying, from the target. Since no particular reason was found for the deviations, these results were not removed from the final calculations. If, however, the results were removed from the dataset, the SD_{bl} became 10.7 U/L, and the difference between the recalculated mean value (176.4 U/L) and the target was still not significant.

With the exception of Dade, the SD_{bl} max criterion was violated by all companies, thus showing a wide dispersion of AST results. Since bias was present and/or SD_{bl} were

Table 1
Mean values, between- and within-laboratory SDs, and maximum allowable between- and within-laboratory SDs (U/L) for the evaluated enzymes

	Mean	SD_{bl}	SD_{bl} max	SD_{wl}	SD_{wl} max		Mean	SD_{bl}	SD_{bl} max	SD_{wl}	SD_{wl} max
ALT						*AST*					
Target	108.3		6.6[1]		11.1[2]	Target	184.0		5.1[1]		9.2[2]
Abbott	105.6	6.4	6.6	1.1	11.3	Abbott	178.4	13.4^{+}	5.1	2.0$^{\#}$	0.9
Beckman	112.9*	6.8	5.1	2.0	9.7	Beckman	177.6	13.9^{+}	5.1	2.8$^{\#}$	0.4
Dade	109.5	5.3	6.6	0.8	12.4	Dade	178.6*	2.7	2.7	1.3	9.2
Olympus	107.3	4.9	6.6	1.2	12.8	Olympus	187.8*	5.2^{+}	3.7	1.8	8.0
Ortho	112.7*	3.4	5.2	1.8	12.6	Ortho	185.7	9.9^{+}	5.1	2.5	4.4
Roche	108.7	6.8	6.6	1.1	11.0	Roche	193.5*	6.5^{+}	0.3	1.7	3.9
CK						*GGT*					
Target	302.7		17.8[1]		29.0[2]	Target	155.3		8.6[1]		9.0[2]
Abbott	306.2*	5.0	18.9	1.4	40.8	Abbott	154.1	10.2	8.6	1.6	7.4
Beckman	312.5*	13.1	15.1	2.5	30.7	Beckman	165.1*	8.8^{+}	4.3	1.8	5.2
Dade	274.2*	10.6^{+}	3.8	2.1	23.4	Dade	160.9	14.1^{+}	8.6	0.7	3.5
Olympus	318.0*	18.6^{+}	11.8	2.9	23.3	Olympus	160.7*	5.5	6.9	1.6	10.2
Ortho	281.5*	15.4^{+}	8.2	6.0	23.0	Ortho	140.1*	5.0^{+}	1.0	0.9	5.6
Roche	288.0*	7.7	12.2	1.9	32.8	Roche	144.0*	3.8	3.3	1.5	8.6
LD						*Amylase*					
Target	351.3		7.7[1]		13.0[2]	Target	70.5		2.8[1]		2.9[2]
Abbott	331.2*	15.0^{+}	0.0	2.1$^{\#}$	0.0	Abbott	62.2*	1.1^{+}	0.0	0.7$^{\#}$	0.5
Beckman	379.5*	11.6^{+}	0.0	2.9$^{\#}$	0.0	Beckman	66.5	7.0^{+}	2.8	0.9$^{\#}$	0.0
Dade	314.4*	18.7^{+}	0.0	1.4$^{\#}$	0.0	Dade	58.7*	1.7^{+}	0.0	0.5$^{\#}$	0.0
Olympus	356.4*	12.6^{+}	7.7	4.9	13.0	Olympus	67.9	7.7^{+}	2.8	0.7$^{\#}$	0.0
Ortho	314.8*	8.2^{+}	0.0	4.3$^{\#}$	0.0	Ortho	53.5*	4.2^{+}	0.0	2.1$^{\#}$	0.0
Roche	347.9*	11.1^{+}	7.7	3.0	13.0	Roche	68.6*	1.3	2.2	0.8	3.6

SD_{bl}, between-laboratory SD; SD_{bl} max, maximum allowable SD_{bl}; SD_{wl}, within-laboratory SD; SD_{wl} max, maximum allowable SD_{wl}.
[1]SD_{bl} max in absence of significant bias; [2]SD_{wl} max in absence of significant bias and presence of SD_{bl} max.
*Significantly different from the target value ($P<0.05$); $^{+}$significantly different from SD_{bl} max ($P<0.05$); $^{\#}$significantly different from SD_{wl} max ($P<0.05$).

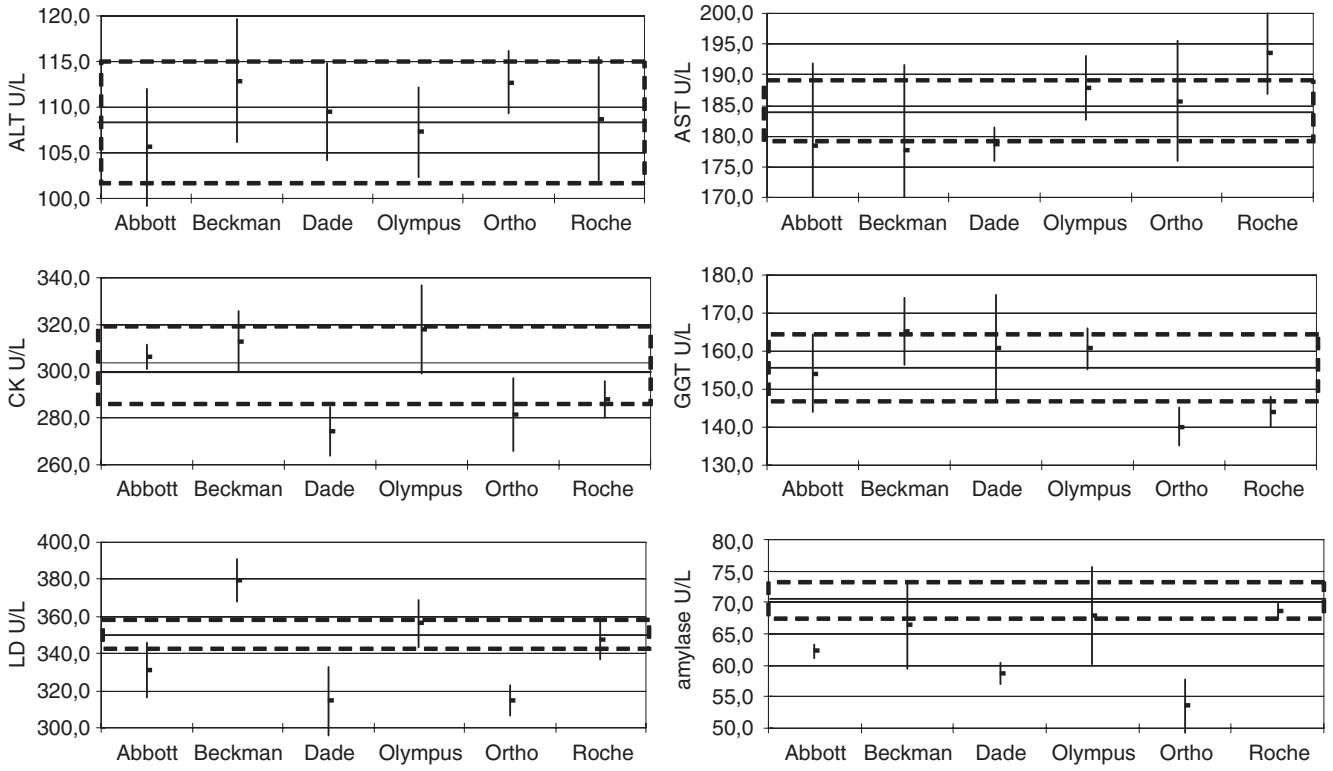

Fig. 1. Target value (fat line), means ± SD_{bl} (U/L) for each company system, and the area (dashed) of maximum allowable SD_{bl} in absence of significant bias.

larger than SD_{bl} max, the SD_{wl} max was smaller than 9.2 U/L for all companies, except for Dade. Nevertheless, for Dade, Olympus, Ortho and Roche systems, the SD_{wl} was well within the corresponding limit. For Abbott and Beckman systems, the total error budget was consumed by the SD_{bl}, so that the corresponding SD_{wl} were significantly larger than the SD_{wl} max.

Except for Dade none of the company assays fully comply with the criteria set (Fig. 1). Based on the study results, it is to be expected that 95% of the laboratories applying Dade systems, 88% of the laboratories applying Olympus and 68% of users of Ortho systems can measure AST within the requirements derived from the biological variation model. 52% of laboratories applying Roche, 50% using Abbott, and 48% Beckman, are expected to measure AST within these limits.

For CK, the biological model allows for relatively large SD_{bl} and SD_{wl}. For a single laboratory, at the target concentration of 302.7 U/L, AB is ±34.8 U/L, SD_d <34.5 U/L, and TAE ± 91.7 U/L.

The mean values per company showed significant bias for all companies (Table 1). The SD_{bl} was larger than the allowable variation for Dade, Olympus, and Ortho systems. The SD_{wl} was much smaller than the SD_{wl} max for all companies.

Despite the significant bias of the mean result for Abbott, Beckman and Roche systems, it is expected that >95% of laboratories applying procedures of these manufacturers will measure CK within the limits of the biological variation model (Fig. 1). For the other three manufacturers, 85% of laboratories using Olympus, 81% using Ortho, and 73% of the laboratories applying Dade procedures are expected to fulfil established criteria.

For GGT, performance criteria derived from the biological variation model for individual laboratories at the target GGT value of 155.3 U/L were ±16.8 U/L for AB, <10.7 U/L for SD_d, and ±34.5 U/L for TAE. A statistically significant bias was shown for Beckman, Olympus, Ortho, and Roche systems. The SD_{bl}s violated the allowable SD_{bl} max criterion for Beckman, Dade and Ortho assays. The SD_{wl} was smaller than the allowable SD_{wl} max for all companies.

98% of laboratories applying Olympus are expected to measure within the limits of the criteria set (Fig. 1). From the study results, it is expected that 92% of the laboratories using analytical procedures of Roche, 90% using Abbott, 79% using Beckman, 73% using Dade, 62% using Ortho will measure GGT within the biological variance model-based limits.

The biologically based performance parameters for individual laboratories measuring LD activities at *T* of 351.3 U/L were ±15.1 U/L for AB, <15.1 U/L for SD_d, and ±40.5U/L for TAE, respectively. Table 1 shows that the means values per company were significantly different from the target for Abbott, Beckman, Dade, and Ortho systems. The bias for these systems was so large that it spent the total error budget and no further variation was allowed nor between- nor within-laboratories. Although without signif-

icant bias, both Olympus and Roche systems violated the allowable SD_{bl} max criterion, while SD_{wl} for these two companies was not statistically larger than the allowable SD_{wl} max.

Fig. 1 shows that the results of none of the companies are within acceptable limits. None of the laboratories using analytical procedures of Ortho are expected to measure LD within the biological variation model limits. For Roche, for which the results of one laboratory were preliminarily excluded, 81% of laboratories will measure within the limits. For Olympus the expected percentage is 73%, for Abbott 37% and finally for Beckman and Dade only 12%.

With regard to α-amylase, the IFCC reference measurement procedure, although available on request, is not yet definitively published. Consequently, the results obtained with the commercially available methods could not be traceable to the IFCC reference system by the time of the present study simply because manufacturers may still not have implemented traceability to this system. Using the *T* of 70.5 U/L, the biological variability model related performance parameters for individual laboratories were ±5.5 U/L for AB, <3.4 U/L for SD_d, and ±11.1 U/L for TAE, respectively. The means per company showed significant negative bias for Abbott, Dade, Ortho, and Roche systems (Table 1). The use of different substrates by some of these manufacturers (and the consequent difference in the analytical specificity of the assays) may in part explain the observed biases. In fact, Abbott and Dade use the chloro-*p*-nitrophenyl-α-D-maltotrioside (CNP-G3) as a substrate and Ortho a dyed amylopectine. On the other hand, Roche uses the same substrate that is used in the IFCC reference measurement procedure, i.e. the ethylidene-4-nitrophenyl-maltoheptaose. In this case, the significance of bias can be explained by the very low SD_{bl} for the group of laboratories using the α-amylase method from this company.

SD_{bl}s were significantly larger than the allowable SD_{bl} max for all company assays, except for Roche. For Dade and Ortho systems, the bias consumed the total error budget and both the allowable SD_{bl} max and SD_{wl} max were zero. For Abbott assay, the allowable SD_{bl} max was zero, but a small SD_{wl} was still allowed. The observed SD_{wl} was, however, significantly larger than the SD_{wl} max. For Beckman and Olympus assays, SD_{bl} was markedly larger than SD_{bl} max leaving a SD_{wl} max of zero.

Fig. 1 shows that Abbott, Beckman, Dade, Olympus, and Ortho systems clearly violated the criteria for amylase measurement. Since for Dade and Ortho the total error budget is used for SD_{bl} and SD_{wl}, the probability for measuring within the established criteria was 0%. Only 1% of laboratories applying Abbott are expected to measure within the limits. For Beckman and Olympus the percentages are 49% and 50%. Though showing a statistically significant bias, the Roche method fulfilled the other criteria, so that >95% of laboratories using this assay are expected to measure amylase within the limits of the biological variation model.

In conclusion, this study shows that the implementation of traceability in the field of clinical enzymology is still not sufficiently realised to reach the expected harmonisation of the values obtained in clinical routine laboratories. Significant bias for the results obtained by commercial analytical procedures was shown in many cases. Furthermore, when desirable analytical performance criteria based on the biological variation model were applied, inappropriate large between- and, in some cases, within-laboratory variances were noticed.

Discussion with the manufacturers and national and international professional groups is needed about the obtained results and the need for future studies. Overall, it appears clear that method bias should be reduced by better calibration to the internationally accepted reference systems. In addition, commercial assays using methodological principles that differ in analytical specificity when compared with the internationally recommended reference systems must be replaced by analytical procedures in which the traceability of calibration to the corresponding IFCC primary reference measurement procedure has been proven experimentally.

Acknowledgement

The authors gratefully acknowledge the co-operation of 70 laboratories from Germany, Italy and the Netherlands.

References

[1] Lex EU. Directive 98/79 EC on in vitro medical devices. Off J L 1998;331:1–37.
[2] Dati F. The new European directive on in vitro diagnostics. Clin Chem Lab Med 2003;41:1289–98.
[3] Müller MM. Implementation of reference systems in laboratory medicine. Clin Chem 2000;46:1907–9.
[4] ISO 18153. In vitro diagnostic medical devices—measurement of quantities in samples of biological origin—metrological traceability of values for catalytic concentrations of enzymes assigned to calibrators and control materials. Geneva, Switzerland: ISO; 2003.
[5] Panteghini M, Ceriotti F, Schumann G, Siekmann L. Establishing a reference system in clinical enzymology. Clin Chem Lab Med 2001;39:795–800.
[6] Schumann G, Bonora R, Ceriotti F, et al. International Federation of Clinical Chemistry and Laboratory Medicine. IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C. International Federation of Clinical Chemistry and Laboratory Medicine: Part 4. Reference procedure for the measurement of catalytic concentration of alanine aminotransferase. Clin Chem Lab Med 2002;40:718–24.
[7] Schumann G, Bonora R, Ceriotti F, et al. International Federation of Clinical Chemistry and Laboratory Medicine. IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C. International Federation of Clinical Chemistry and Laboratory Medicine: Part 5. Reference procedure for the measurement of catalytic concentration of aspartate aminotransferase. Clin Chem Lab Med 2002;40:725–33.
[8] Schumann G, Bonora R, Ceriotti F, et al. IFCC primary reference procedures for the measurement of catalytic activity concentrations of

enzymes at 37 °C: Part 2. Reference procedure for the measurement of catalytic concentration of creatine kinase. Clin Chem Lab Med 2002;40:635–42.
[9] Schumann G, Bonora R, Ceriotti F, et al. International Federation of Clinical Chemistry and Laboratory Medicine. IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C. International Federation of Clinical Chemistry and Laboratory Medicine: Part 6. Reference procedure for the measurement of catalytic concentration of gamma-glutamyltransferase. Clin Chem Lab Med 2002;40:734–8.
[10] Schumann G, Bonora R, Ceriotti F, et al. IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37 °C: Part 3. Reference procedure for the measurement of catalytic concentration of lactate dehydrogenase. Clin Chem Lab Med 2002;40:643–8.
[11] Jansen RTP, Jansen AP. Standards versus standardized methods in enzyme assay. Ann Clin Biochem 1983;20:52–9.
[12] Thienpont LM, Stöckl D, Friedecký B, Kratochvíla J, Budina M. Trueness verification in European external quality assessment schemes: time to care about the quality of the samples. Scan J Clin Lab Invest 2003;63:195–202.
[13] Thienpont LM, Stöckl D, Kratochvíla J, Friedecký B, Budina M. Pilot external quality assessment survey for post-market vigilance of in vitro diagnostic medical devices and investigation of trueness of participants' results. CLin Chem Lab Med 2003;41:183–6.
[14] Jansen RTP. Kalibratie 2000. Ned Tijdschr Klin Chem 1998;23: 261–4.
[15] Jansen RTP. Calibration 2000: state of the art, relation with IVD-directive, future. Ned Tijdschr Klin Chem Labgeneesk 2005;30: 49–55.
[16] Baadenhuijsen H, Steigstra H, Cobbaert C, Kuypers A, Weykamp C, Jansen R. Commutability assessment of potential reference materials using a multicenter split-patient-sample between-field-methods (Twin-Study) design: study within the framework of the Dutch project "Calibration 2000". Clin Chem 2002;48:1520–5.
[17] Cobbaert C, Weykamp C, Baadenhuijsen H, Kuypers A, Lindemans J, Jansen R. Selection, preparation, and characterization of commutable frozen human serum pools as potential secondary reference materials for lipid and apolipoprotein measurements: study within the framework of the Dutch project "Calibration 2000". Clin Chem 2002;48:1526–38.
[18] Baadenhuijsen H, Kuypers A, Weykamp C, Cobbaert C, Jansen R. External quality assessment in The Netherlands: time to introduce commutable survey specimens. Lessons from the Dutch "Calibration 2000" project. Clin Chem Lab Med 2005;43:304–7.
[19] Fraser CG. General strategies to set quality specifications for reliability performance characteristics. Scand J Clin Lab Invest 1999; 59:487–90.
[20] Ricos C, Alvarez V, Cava F, et al. Current databases on biologic variation: pros, cons and progress. Scand J Clin Lab Invest 1999;59: 491–500.
[21] Buhl SN, Jackson KY, Graffunder B. Optimal reaction conditions for assaying human lactate dehydrogenase pyruvate to lactate at 20, 30, and 37 °C. Clin Chem 1978;24:261–6.