DE GRUYTER
Clin Chem Lab Med 2018; 56(10): 1673–1684

Review

Rob T.P. Jansen*, Christa M. Cobbaert, Cas Weykamp and Marc Thelen

The quest for equivalence of test results: the pilgrimage of the Dutch Calibration 2.000 program for metrological traceability

https://doi.org/10.1515/cclm-2017-0796
Received September 4, 2017; accepted November 17, 2017; previously published online January 17, 2018

Abstract: Calibration 2.000 was initiated 20 years ago for standardization and harmonization of medical tests. The program also intended to evaluate adequate implementation of the In Vitro Diagnostics (IVD) 98/79/EC directive, in order to ensure that medical tests are fit-for-clinical purpose. The Calibration 2.000 initiative led to ongoing verification of test standardization and harmonization in the Netherlands using commutable external quality assessment (EQA)-tools and a type 1 EQA-design, where feasible. National support was guaranteed by involving all laboratory professionals as well as laboratory technicians responsible for EQA and quality officers. A category 1 EQA-system for general chemistry analytes, harmonizers for specific analytes like hGH and IGF-1, and commutable materials for other EQA-sections have been developed and structurally introduced in the EQA-schemes. The type 1 EQA-design facilitates the dialogue between individual specialists in laboratory medicine and the IVD-industry to reduce lot-to-lot variation and to improve standardization. In such a way, Calibration 2.000 sheds light on the metrological traceability challenges that we are facing and helps the laboratory community to get the issues on the table and resolved. The need for commutable trueness verifiers and/or harmonizers for other medical tests is now seen as paramount. Much knowledge is present in the Netherlands and for general chemistry, humoral immunology and protein chemistry, a few endocrinology tests, and various therapeutic drug monitoring (TDM) tests, commutable materials are available. Also the multi sample evaluation scoring system (MUSE) and the category 1 EQA-design offer many possibilities for permanent education of laboratory professionals to further improve the between and within laboratory variation and the test equivalence.

Keywords: equivalence; harmonization; quality; standardization; traceability.

***Corresponding author: Rob T.P. Jansen,** SKML, Mercator 1, Toernooiveld 214, 6525EC Nijmegen, The Netherlands, E-mail: rtpjansen@gmail.com
Christa M. Cobbaert: LUMC, Department of Clinical Chemistry, Leiden, The Netherlands
Cas Weykamp: Queen Beatrix Hospital, MCA Laboratory, Winterswijk, The Netherlands
Marc Thelen: Amphia Hospital, Clinical Chemistry and Haematology, Breda, The Netherlands

Introduction

Medical diagnosis and treatment should be comparable for patients across physicians, health care centers and countries. Medical test results are paramount in diagnosis and monitoring of treatment in a large majority of medical decisions. According to the In Vitro Diagnostics (IVD) directive 98/79/EC and ISO 17511:2003 medical test results should be traceable to standards of higher order in order to allow equivalency of test results between instruments as well as between laboratories and countries. Equivalence can be defined as agreement among results of a laboratory test within the clinically meaningful limits. Clinically meaningful limits are defined as the maximum degree of variability in results of laboratory tests that allows optimal patient care.

Equivalence of laboratory data is important not only within laboratories between instruments, but also in time, when monitoring patients. As we live in a global world, equivalence in reported concentration levels and in units used is needed for unequivocal diagnosis and medical decision making. In addition when patients are transferred from one center to another, equivalence of data is essential. Test imprecision and bias both should fulfill predefined analytical performance criteria derived from either clinical outcome or biological variation, to guarantee that the test is fit-for-clinical purpose [1, 2]. Units should adhere to the BIPM/IFCC agreed International System (SI) of units. Comparable diagnoses and treatments require equivalence of laboratory data. To that end

bias, imprecision and tolerance criteria should be based on agreed models for analytical performance to enable better patient outcome.

Equivalence is needed in all disciplines of laboratory medicine, including general chemistry, endocrinology, immunology, hematology, coagulation, therapeutic drug monitoring (TDM), virology and others.

The Calibration 2000 program was initiated in 1998 by Rob Jansen, at the 25th anniversary of SKML, the Dutch external quality assessment (EQA)-organization, now 20 years ago [3, 4]. The program aimed at national standardization of medical tests (first choice) whenever feasible, and at harmonization (second best) in case that standardization was not feasible. Standardization is the situation when patient results are equivalent between measurement procedures and calibration is traceable to SI by use of a reference measurement procedure. Harmonization can be defined as a process that reduces the variability of results of a laboratory measurement procedure to a level within the clinically meaningful limits. Six SKML EQA-sections were involved from the beginning. It started as a national initiative aiming at equivalence of laboratory test results [5, 6]. Fifteen years later the program was renamed Calibration 2.000 to emphasize that the program entered a second phase after achieving its first successes, without loosing its original ambition and aiming at standardization or, if not possible, harmonization of tests at the global level. The Calibration 2.000 program that has been developed through the past two decades is summarized in Figure 1 and encompasses a unified, national EQA-approach across sections based on a nationally developed toolbox at MCA Winterswijk, an ISO 13485 accredited EQA-material production center, to reach equivalence of medical test results. It includes three principal components:

1. Development of value-assigned, commutable EQA-materials, covering clinically relevant concentration ranges, structurally embedded in the national EQA-surveys in combination with a tolerance limit system based on the Stockholm conference hierarchy (previously) and the Milan hierarchy (currently) for analytical performance specifications. These EQA-materials are considered as “holy grail” tools and are used as trueness verifiers in the case of SI-traceable tests, to measure unequivocally the analytical performance of SI-standardized tests.
2. Development of a fair scoring system to analyze conformity of tests to preset analytical performance criteria. The scoring system comprises a multi sample evaluation (MUSE) EQA system, used in combination with trueness verifiers (see 1).

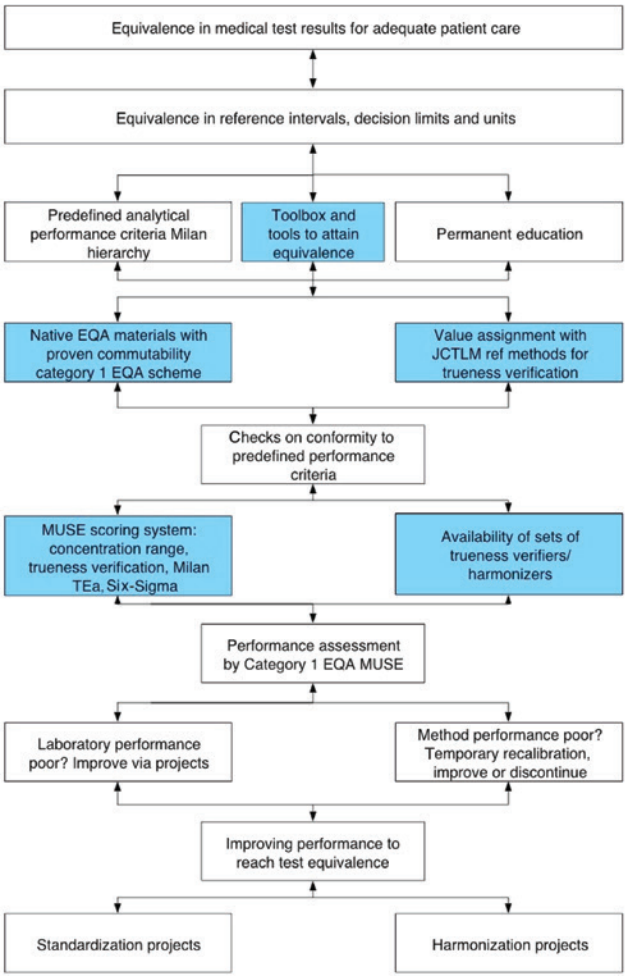

Figure 1: Calibration 2.000 approach to reach equivalence of medical test results.
The essential parts of the toolbox are color coded. All items are covered by Calibration 2.000. Milan hierarchy from [7]; MUSE scoring system from [8]. EQA, external quality assessment; JCTLM, Joint Committee on Traceability in Laboratory Medicine; MUSE, multi sample evaluation; TEa, total error allowable. Harmonization projects conform [9].

3. Prioritized standardization/harmonization projects to maintain good quality and to improve poor ones. The projects included standardization where attainable, and harmonization in all other situations.

Pre-requisites for Calibration 2.000

The approaches used were comparable ‘avant la lettre’ to the integrated harmonization protocol described in the toolbox of technical procedures for harmonization at http://www.harmonization.net/media/1004/tool_box_2013.pdf.

Standardization and harmonization projects were prioritized based on unmet clinical needs and feasibility. The ultimate goal of Calibration 2.000 is to contribute to better patient care by delivering the right test result for the right patient at the right time. Accurate results are key for adequate patient care in which laboratory data play a major role, which is the case in the majority of medical decisions (70% rule). To that end, metrological traceability of laboratory data according to ISO 17511:2003 is pertinent.

For correct interpretation of test results, matching clinical reference intervals and decision limits are key in combination with SI-units whenever feasible. Absolute concentrations are often used in clinical decision making (e.g. cholesterol, creatinine or derived parameters, thyroid stimulating hormone, glucose) whereas interlaboratory and intralaboratory differences, especially differences in analytical set-points, can markedly affect the clinical interpretation of tests [10]. This also holds for reference intervals. The use of standardized reference intervals, decision limits and SI-units is still debated at the international level but currently gets attention from European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working groups. However, standardization of units should not be controversial as the use of different units for results of the same analyte can lead to clinical misinterpretation and patient harm [11]. In the Netherlands, the discussion on the necessity of molar standardization started in the 1960s and led at that time to a controversy between internal medicine doctors and specialists in laboratory medicine, described in the Blind Mole [12]. Since then, Dutch medical laboratories are nationwide using SI-units.

Definition of preset analytical performance criteria for the needed level of equivalence were obtained at first from the Stockholm conference on strategies to set global quality specifications in laboratory medicine [13], and more recently from the 1st Strategic Conference of the EFLM on defining analytical performance specifications [7]. The Milan conference criteria based on clinical outcome studies (Model 1) and biological variation (Model 2) were the preferred models. State of the art based criteria were considered second best.

Tools developed for the Calibration 2.000 program were commutable trueness verifiers or harmonizers and a targeting system. Standardization or harmonization can only be reached when targeted commutable materials are available for trueness verification. The development of such materials was considered as the 'Holy Grail' of the Calibration 2.000 program. As well, a toolbox of technical procedures to be considered for developing a process to achieve harmonization or standardization of measurands formed the basis.

In the program, participating reference laboratories for enzymes, lipids, PT/INR and HbA_{1c} are encouraged to apply for ISO 15195 and ISO 17025 accreditation in order to get their reference measurement procedures internationally recognized and Joint Committee for Traceability in Laboratory Medicine (JCTLM)-listed. The development of new candidate reference methods is also recommended.

Calibration 2.000 is a national endeavor and is progressed by joint effort and collaboration among different laboratory specialisms after creating national support. Dutch laboratory specialists are excellent organizers, firm in their principles, consensus oriented and scientifically engaged. The EQA-sections involved in standardization/harmonization are all occupied by volunteering laboratory professionals who want to push standardization and harmonization forward, for the sake of better patient care with the focus being on quality. Also, medical laboratories are mostly led by professionals who are specialists in laboratory medicine and carry end-responsibility for the services of medical laboratories.

Education is a major aspect of Calibration 2.000 focused on specialists in laboratory medicine and medical laboratories. All SKML sections have closed quarterly meetings but share expertise during yearly conferences attended by specialists in laboratory medicine and also by technicians and quality officers involved with EQA. The Calibration 2.000 steering committee has biannual meetings which provide a podium for aligning standardization efforts and sharing expertise. Education consists of lectures, courses, and symposia on topics such as standardization, harmonization, method standardization, validation procedures and explanation of EQA results. Also the lay out and parameterization of the EQA system contributes to education, e.g. the method definition for enzymes is restricted to two method groups, for methods that are either IFCC traceable or not. This gives laboratory professionals insight and triggers them to opt for IFCC-traceability.

Toolbox

Tools to reach equivalence are (1) commutable trueness verifiers and harmonizers, (2) value assignment and (3) checks.

1. Commutable materials were produced from fresh frozen patient samples. Commutability was assessed originally in a twin study design [14] and more recently according to CLSI EP 30A (formerly C53-A) [15].
2. Value assignment of the materials was done according to the traceability chain using internationally

recognized reference methods, reference materials and reference laboratories. Bias and imprecision criteria are derived from the Milan conference models. In addition the Six Sigma metric was introduced as a measure of allowable error.

3. Checks to analyze conformity to the preset criteria were constituted in two systems, (1) an ongoing national EQA scheme using multi sample evaluation, and (2) provision of trueness verification sample sets for individual laboratories for local analytical performance checks.

With the above described approach and toolbox for standardization/harmonization of measurands, the Dutch EQA scheme organization SKML established a Category 1 EQA system [8, 16], according to the criteria described in the 'Roadmap to Harmonization' paper [17, 18]. According to the definition of a Category 1 EQA scheme [18] the SKML scheme uses commutable native sera with value assignment by reference methods and allows for verification of trueness and precision of IVD tests based on predefined tolerance limits of measurement errors to evaluate whether medical tests are fit-for-purpose [19]. It incorporates the following:

1. Trueness verification materials are offered to laboratories for validating/verifying new or reset methods or instruments.
2. Analytical performance is assessed by the MUSE scoring system [8]. The system assesses both the performance of participating laboratories and their methods. Poor (and adequate) performance can result from poor (and adequate) laboratory performance or from the method performance. For example, the Jaffe methods for creatinine perform poorly and should be replaced for more adequate performance of the laboratory [20, 21]. Also the suitability of the analytical performance of current tests in medical laboratories, in the light of guidelines for diabetes and ischemic heart disease, was questioned [22].
3. Improvement of trueness and imprecision was and is driven by successive standardization projects, where attainable, or otherwise harmonization projects both nationally and internationally.

The Calibration 2.000 activities catalyzed the transformation of the Dutch SKML from a traditional EQA-organization with a peer group method approach to an EQA-organization that embedded the metrological traceability and commutability concepts, in line with upcoming IVD regulations, ISO-guidelines and international standardization efforts such as those coordinated by JCTLM and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) in the last 20 years. Based on this national and holistic approach, in combination with a well thought out toolbox and EQA-design for harmonization of measurands, the current SKML-EQAS gives specialists in laboratory medicine actionable information about test performance. The toolbox and Calibration 2.000 approach to achieve equivalence of test results are summarized in Figure 1. Most items were set up in the program itself. In Figure 1 focus is on analytes for which reference methods and/or reference materials are available. For analytes where these are not in place steps towards harmonization are being set in line with ISO 17511 (under development) in which a special harmonization chain will be described. Presently in the EQA schemes expert values are being used as target values rather than consensus values. Where no such target value is available scoring is omitted in the scheme.

Contributors to greater harmonization

Equivalence of laboratory data is pertinent for accurate diagnosis and treatment of patients. The perception of patients and medical doctors regarding medical laboratory services is that diagnostic laboratories deliver accurate test results, which guarantee an unequivocal and comparable risk classification, diagnosis, treatment and patient outcome across hospitals. External quality assurance programs, which structurally evaluate laboratory test performance, demonstrate that this perception is questionable if not wrong, even for routine medical tests [23].

Standardization and harmonization are hot topics in laboratory medicine. Many papers address the issue and several initiatives were taken to improve the differences seen. In particular the AACC harmonization initiative [17, 18], initiatives of IFCC and EFLM [24–28] and the installation of the JCTLM contributed significantly, as did the Calibration 2.000 project.

The issue of allowable error has gained much interest. The Stockholm conference and later on the Milan conference contributed much to awareness and agreement on the way forward. Clinical outcome studies and particularly biological variation based estimations of allowable error are widely accepted [1, 7, 13, 27–29].

The need for sophisticated EQA was clearly indicated [18]. In Calibration 2.000 a Category 1 EQA scheme was developed using commutable samples that were value assigned with reference methods and covering

the concentration range of interest, and using biological variation based trueness and imprecision tolerance limits including Six Sigma metric [8, 16]. The development of commutable samples made it possible to perform national and international studies on standardization and harmonization. Jansen and Jansen [30], Boerma et al. [31] and Jansen et al. [32] showed already in the 1980s and 1990s that harmonization is necessary and standardization should be possible, though the materials used were insufficient and lacked commutability. The development of such materials, the value assignment by reference laboratories and establishment of Category 1 EQA schemes, make it possible at present to really assess whether the metrological traceability concept is implemented in an adequate way so that test standardization/harmonization is achieved. Jansen et al. [33] showed this was not the case even for serum enzymes.

Implementation of harmonization activities across disciplines

It was very encouraging that Calibration 2.000 was embraced by several medical laboratory disciplines. The milestones of the program are summarized in Table 1. The detailed achievements of the various disciplines participating are summarized below.

General chemistry

Most initiatives for harmonization and standardization have been in the field of general chemistry. The development of commutable EQA materials [33, 42, 54, 55], the first 'Holy Grail' materials, made evaluation of interlaboratory differences possible. In the studies it was shown that harmonization was achievable for several analytes, at first for lipids and also for serum enzymes [42, 43, 54, 55]. For lipids in national studies interlaboratory variation was shown in normal lipemic commutable samples [42]. The introduction of biological variation based tolerance limits made objective judgment of the performance of laboratories and methods possible. Wide variation with both acceptable and unacceptable performance was shown in international studies for lipids, enzymes, creatinine, electrolytes and other analytes [16, 33, 45, 56]. The joint project by the national EQA organizers in Italy, the Netherlands, Portugal, UK and Spain (INPUtS) [45] showed that the analytical performance of 11 of 17 general chemistry analytes measured in European medical laboratories met the minimum performance specifications. However, only one analyte (creatine kinase) met the desirable specifications in all countries and for all manufacturers. There were however, major differences between other analytes. There were six analytes for which the minimum quality specifications are not met and manufacturers should strive to improve their performance for these analytes. Standardization of enzyme methods towards IFCC

Table 1: Milestones and deliverables of Calibration 2.000 program from 1998 to 2018.

Year	Milestone	Refs.
1998	Initiation at congress celebrating the 25th anniversary of the Dutch external quality assessment (EQA)-organization SKML	[3–5]
1999	Task Force Calibration 2.000 8 sections of SKML	[6, 34–39]
2001	Twin studies to detect first commutable materials	[14, 40, 41]
2002	First commutable and value-assigned materials for general chemistry	[42, 43]
2005	Nationwide introduction of trueness verification in EQA-scheme, including ongoing commutability assessment with native spy material	[43, 44]
2006	International enzyme studies	[16, 33, 45]
2008	National harmonization of growth hormone	[34, 46]
2009	National harmonization of fibrinogen, Factor VIIIc, antithrombin	[36, 47, 48]
2010	>60% laboratory standardized for eight general chemistry analytes	[44]
2012	National harmonization of seven enzymes	[44]
2014	Renaming to Calibration 2.000 because of global scope	
2014	Participation in preparational harmonization conference of the American Association for Clinical Chemistry (AACC)	[17, 18]
2015	Introduction of a uniform multi sample evaluation scoring system for all SKML sections	[8, 16]
2015	Commutable material for therapeutic drug monitoring (TDM)	[37, 49]
2017	Development of reference methods/systems for serum apolipoproteins, plasma antithrombin, serum free light chains	[50–53]

reference measurement procedures (RMPs) has not been fully achieved as shown in the INPUtS study. Forty-four percent of results of seven enzymes on six platforms failed the desirable criterion based on biological variation. It requires ongoing efforts. Performance of laboratories and tests showed under-performance in several studies, and the necessity of better standardization of both methodology and calibration materials was stressed [20–23].

HbA_{1c} is another test that needed harmonization [57]. Also for HbA_{1c} important results were achieved. An international HbA_{1c} network has been established for global standardization of HbA_{1c}. In the Netherlands, two laboratories operate the international reference method for HbA_{1c} and are recognized as certified reference laboratories. Practically all Dutch laboratories implemented in a coordinated action with clinical societies and patient organizations the new units of mmol/mol. The improvement in HbA_{1c} test performance is huge as interlaboratory CVs reduced from 30% in the early 1990s to about 3% in 2017. Stressing necessity is one thing, achieving the objectives is another. Figure 2 shows improvements reached in analytical performance in the Netherlands for

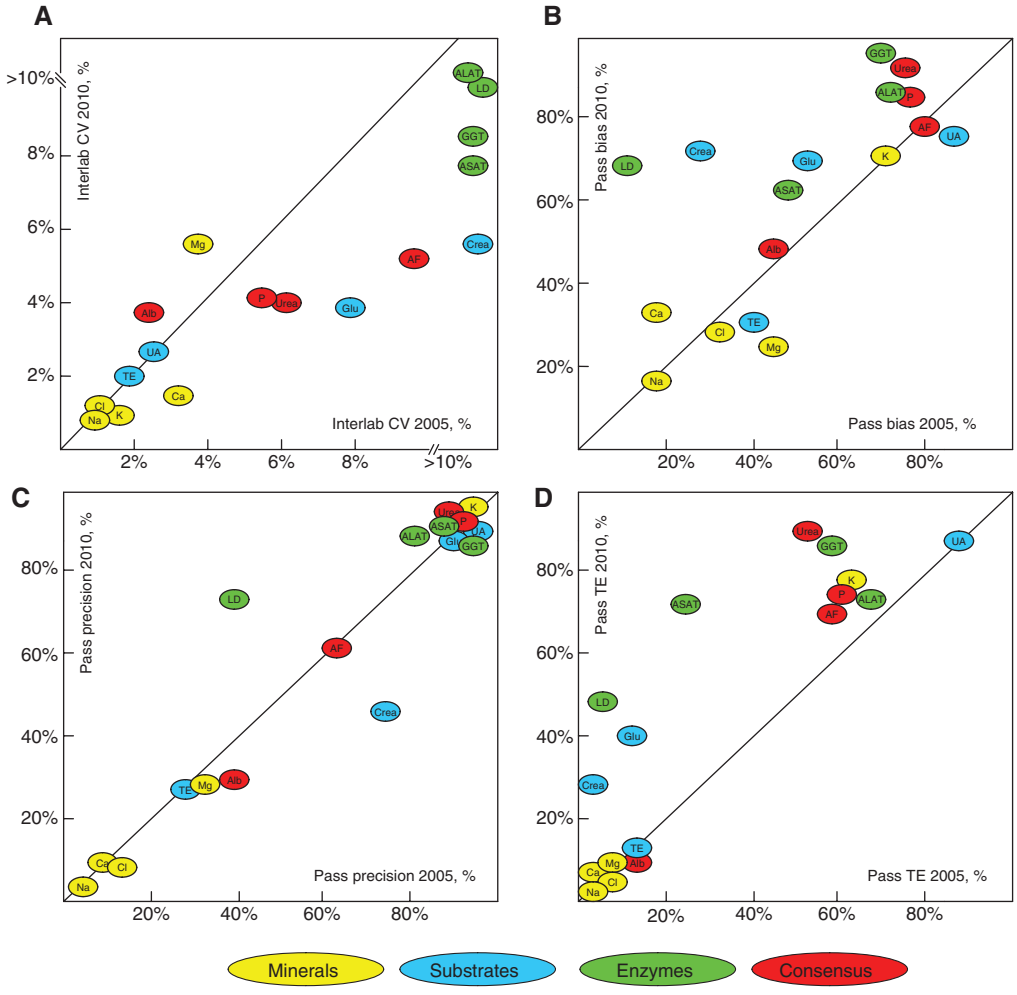

Figure 2: Analytical performance trends for 17 general clinical chemistry parameters between 2005 and 2010.
The analytes are divided in four color coded groups: three color coded groups refer to minerals, substrates and enzymes of which the EQA-materials have been value assigned with recognized reference methods (n = 13); the red colored group encompasses the parameters that rely on consensus means and have not been value assigned with recognized reference methods (n = 4). (A) Evolution of the degree of equivalence of test results as indicated by the between-laboratory CV (CV in 2010 is smaller than in 2005 for most analytes). (B) Evolution of the trueness component between 2005 and 2010. Criterion for bias is based on deviation from reference value in relation to the biological variation desirable criterion [1, 2]. The percentage laboratories that pass the bias criterion is larger in 2010 than in 2005. (C) Evolution of the precision component between 2005 and 2010 criterion for imprecision (within-laboratory CV) is based on biological variation desirable criterion [1, 2]. The percentage laboratories that pass the CV criterion is about equal in 2010 and 2005. (D) Evolution of total allowable error (TEa) between 2005 and 2010 Criterion for total error is based on biological variation desirable criterion [1, 2]. The percentage laboratories that pass the TE criterion is larger in 2010 than in 2005. For several parameters only 20% pass the TE criterion because of the rather strict biological variation criteria particularly for sodium, chloride, calcium and magnesium. (Reprinted from [44], with permission from the publisher Elsevier.)

enzymes, creatinine and other analytes after running the type 1 EQA-program for 5 years [44, 58]. Additional challenges were demonstrated when lipemic sera were used for testing lipid methodology [59] or high glucose sera in creatinine analysis [60]. The demonstration of poor performance of non-IFCC traceable methods for enzymes and of Jaffe methods for creatinine led to better standardization of methodology in the Netherlands [44].

Reference laboratories operating internationally recognized reference methods exist in the Netherlands for serum lipids (the Lipid Reference Laboratory of Erasmus University Hospital which is a member of the CRMLN operated by CDC), serum enzymes (Dr. Paul F.H. Franck, LabWest, The Hague) and HbA_{1c} (Dr. Cas Weykamp, Beatrix Hospital Winterswijk, Dr. Erna Lenters, Isala Hospital Zwolle).

Lipid and apolipoprotein standardization

Notwithstanding adequate standardization of direct high-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C) tests, it was demonstrated at the national level that direct HDL-C and LDL-C tests suffer from non-selectivity in hypertriglyceridemic sera, leading to significant misclassification rates [59]. As the residual cardiovascular risk after reaching the treatment goals for LDL-C, blood pressure and diabetes remains high (around 70%), alternative (apolipo)protein tests should help to unravel cardiovascular disease pathophysiology and to select better treatment targets. To that end, an IFCC working group on apolipoprotein (APO) standardization has been established which should develop a complete mass spectrometry based Reference Measurement System for multiplex apolipoprotein quantitation and phenotyping (http://www.ifcc.org/ifcc-scientific-division/sd-working-groups/wg-apo-ms/). Investigations have been started to metrological traceability in mass spectrometry-based targeted protein quantitation of the apolipoproteins A-I and B100 [61].

Endocrinology, binding assays, vitamins

The issue of harmonization and standardization was also recognized in endocrinology and binding assays as tumor markers and vitamins. Although reference methods exist in the JCTLM database for several steroid hormones, mostly based on mass-spectrometry methods, standardization is still a major problem. There is a lack of commutable materials in EQA surveys to check routine tests which are often based on immunological methods.

Peptide hormones are even a bigger challenge. A special problem in this respect are growth hormone (hGH) and insulin like growth factor (IGF-1). The diagnosis growth hormone deficiency largely depends on the hGH test result. Because of the relevance of growth hormone treatment in children with short statue, both clinically and financially, between-laboratory harmonization of results is paramount. A practical approach in the Netherlands led to such national harmonization using a national harmonizer [34, 46, 62]. This initiative is now taken to the international level in an international IFCC working group on hGH standardization, led by Eef Lentjes from the Netherlands (http://www.ifcc.org/ifcc-scientific-division/sd-working-groups/growth-hormone-(wg-gh)/).

Today carbohydrate deficient transferrin (CDT) is considered an important objective indicator of sustained high alcohol consumption and is used to support the diagnosis of alcohol abuse and dependence in medical and forensic settings, including driver's license withdrawal and reinstatement [63]. Harmonization of the assay obviously is important. The Calibration 2.000 project and co-workers of it contributed both to harmonization [64, 65] and to the development of a reference method [63].

Another example of an analyte in need of harmonization is vitamin B6 (pyridoxal/pyridoxal-phosphate). van Zelst et al. [35] showed large within-laboratory variation for some methods, and in addition between-laboratory variation. The authors suggested that the lack of a reference method or suitable certified reference material for the measurement of vitamin B6 in whole blood is impeding the standardization or harmonization of this assay.

Protein chemistry

In various fields of protein chemistry studies were done towards harmonization. Initial studies to find commutable material for serum proteins such as albumin, α_1-antitrypsin, immunoglobulins and other proteins were done in the Netherlands by Klasen et al. [40, 66]. Successful follow-up of the initiative has found application with commutable materials, value assigned in the SKML EQA system for plasma proteins, but this has not yet been published. For other more specific proteins research focuses on method robustness and harmonization. For example, first steps are underway for hepcidin [67].

The use of NIST SRM 2921 and recombinant cTnI-based serum pools for harmonization was investigated by Cobbaert et al. [50, 68] who found the material unstable.

Coagulation

Harmonization of coagulation data is important in particular for situations including monitoring anticoagulation therapy such as prothrombin time INR, diagnosis and monitoring of factor VIII-C, for decision levels of fibrinogen, antithrombin and activated partial thromboplastin time. The SKML section on coagulation participated actively in the Calibration 2.000 project from the start and harmonization initiatives were taken for fibrinogen, Factor VIIIc and antithrombin [36, 41, 47, 48, 69]. Extensive scientific work on PT/INR standardization was done at the Leiden reference laboratory (Dr. Ton van den Besselaar), which runs the reference method for INR. PT/INR standardization activities are currently embedded in the Coagulation Reference Laboratory of the Department of Clinical Chemistry and Laboratory Medicine, at LUMC, which has initiated global standardization of the Manual Tilt Tube method for PT/INR standardization. Also, a collaboration between the Coagulation Reference Lab in Leiden, the international EQA organization for coagulation (ECAT) and IVD-industry has been set up for developing a mass spectrometry based reference measurement system for antithrombin [51].

Therapeutic drug monitoring

Recently the Calibration 2.000 tools and toolbox were also applied for standardization of TDM tests. Commutable materials were developed for carbamazepine, valproic acid and tobramycin [37, 49]. Between-laboratory studies showed that some methods were inaccurate. Future studies will indicate whether the use of the commutable materials could lead to harmonization between laboratories and methods.

Hemocytometry and flow cytometry

Initial studies indicated that commercially available materials cannot be used as a calibrator for all parameters on all of the common hematology analyzers. Most EQA samples, however, as prepared and used in the Netherlands, were commutable for all parameters and analyzers [70]. Most routine parameters including RBC, WBC, platelets and Hb are harmonized in the Netherlands. Future studies are necessary to show harmonization of the parameters RBC-distribution width and platelet-distribution width and also of WBC subclasses.

For flow cytometry of the parameters HLA-B27, CD-45 and CD-34+ lymphocytes, studies were performed and showed inter-laboratory variation [38, 71–74]. The competition with PCR methodology hampered further efforts.

Microbiology

Cut-off values in the sense of positive or negative for serology parameters could have a clinical impact if there is lack of harmonization. Studies were started in this respect in line with the early onset of Calibration 2.000 [39, 75]. Also in this field further studies are needed to develop commutable materials and promote calibration.

Current Calibration 2.000 projects

Currently the Calibration 2.000 steering committee focuses on sustainable contributions to standardization and harmonization at both the national and international level. Some sections are supporting the development of reference methods for example for standardization of serum apolipoproteins C1, C2, C3 and E [76], serum free light chains [77], and citrate plasma antithrombin [51]. Also, an approach and system for implementing national reference intervals and decision limits for SI-traceable analytes is underway [78]. Finally, as part of continuing education SKML EQA sections will start reporting only in the preferred unit, instead of the unit of choice by the participant.

Lessons learnt from Calibration 2.000

There is worldwide an increasing awareness of the need for test standardization and harmonization in medical practice [52, 53]. Also, international ISO guidelines including ISO 17511, ISO 17025, ISO 15189 and mandatory legislation such as the IVD directive 98/79/EC and upcoming new IVD Regulation in Europe, demand global standardization of medical tests, if feasible. Multiple international and national communities and stakeholders of medical tests have taken initiatives in an attempt to reach test equivalence [3, 17, 52].

Standardization of units and introduction of molar units were already realized in the 1970s in the Netherlands [12]. National support for test standardization since 1998 combined with the striving of SKML and Calibration 2.000 Steering Committee to develop type 1 EQA schemes across several EQA-sections and laboratory disciplines, resulted in the development of a permanent infrastructure and

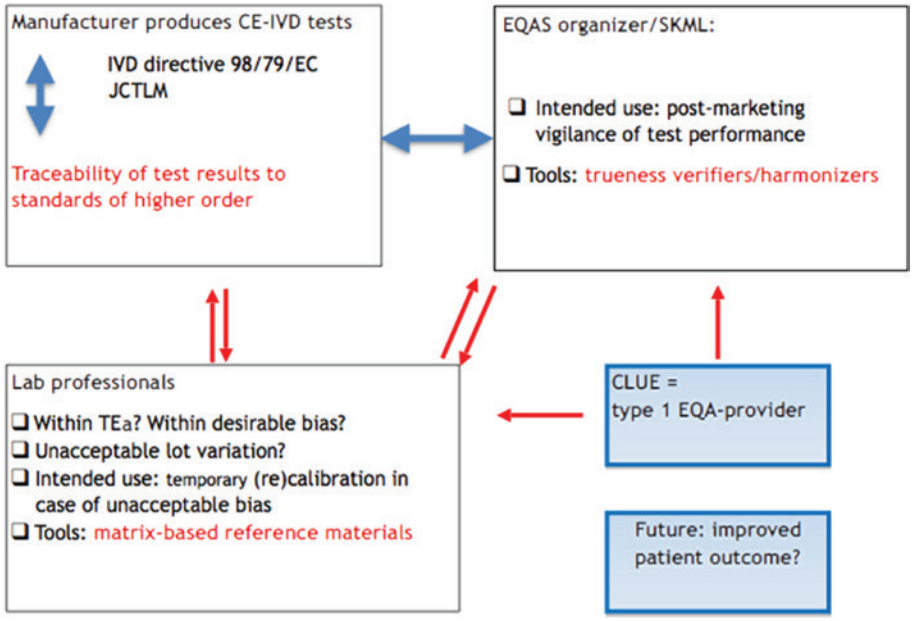

Figure 3: Relationship between IVD-industry, SKML Calibration 2.000 and laboratory professionals.
CE-IVD, Conformité Européenne (European conformity) *in vitro* diagnostics; JCTLM, Joint Committee on Traceability in Laboratory Medicine; EQAS, external quality assessment scheme; TEa, total error allowable.

powerful toolbox for giving insight into trueness (for SI-traceable tests), precision and interlaboratory variation of medical tests. The Dutch EQA-design and toolbox allow laboratory professionals to act by means of, e.g. temporary recalibration if inequality of test results has occurred and may cause patient harm.

Unfortunately, standardization and harmonization of tests is cumbersome for many reasons including regulations that hinder uncomplicated re-standardization and harmonization of existing tests (e.g. FDA required test re-evaluation after re-calibration). On the other hand, notwithstanding successful worldwide realization of for example an IFCC reference method system for HbA_{1c}, adoption of re-standardized tests is slow at the international level as there are still two reference systems (IFCC and NGSP) and two units (mmol/mol and %) in place. The same holds for the standardization of the serum enzymes: there are still IFCC traceable and non IFCC traceable enzyme results.

Calibration 2.000 was initiated 20 years ago for standardization and harmonization of medical tests. The program also intended to evaluate adequate implementation of the IVD 98/79/EC directive, to ensure that medical tests are fit-for-clinical purpose. Figure 3 shows the relationships between the IVD-industry, SKML-Calibration 2.000 and the laboratory professionals. The Calibration 2.000 initiative led to ongoing verification of test standardization and harmonization in the Netherlands using commutable EQA-tools and a type 1 EQA-design, where feasible. National support was guaranteed by involving all laboratory professionals as well as laboratory technicians responsible for EQA and quality officers. A category 1 EQA-system for general chemistry analytes, harmonizers for specific analytes like hGH and IGF-1, and commutable materials for other EQA-sections have been developed and structurally introduced in the EQA-schemes. The type 1 EQA-design facilitates the dialogue between individual specialists in laboratory medicine and the IVD-industry to reduce lot-to-lot variation and to improve standardization. In such a way, Calibration 2.000 sheds light on the metrological traceability challenges that we are facing and helps the lab community identify the issues and to resolve these. The need for commutable trueness verifiers and/or harmonizers for other medical tests is now seen as paramount. Much knowledge is present in the Netherlands and for general chemistry, humoral immunology and protein chemistry, a few endocrinology tests, and various TDM tests, commutable materials are available. Also the MUSE scoring system and the category 1 EQA-design offer many possibilities for permanent education of laboratory professionals to further improve the between and within laboratory variation and the test equivalence.

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

Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Fraser CG. General strategies to set quality specifications for reliability performance characteristics. Scand J Clin Lab Invest 1999;59:487–90.
2. Ricos C, Alvarez V, Cava F, García-Lario JV, Hernández A, Jiménez CV, et al. Current databases on biologic variation: pros, cons and progress. Scand J Clin Lab Invest 1999;59:491–500.
3. Jansen RT. Kalibratie 2000. Ned Tijdschr Klin Chem 1998;23:261–4.
4. Kuypers A, Baadenhuijsen H, Jansen R. Calibration 2000: EQAS produced commutable secondary calibrators for your laboratory. Clin Chem Lab Med 2001;39(Special Suppl.):S1–S448, S68 (Abstract).
5. Jansen RT. The quest for comparability: Calibration 2000. Accredit Qual Assur 2000;5:363–6.
6. Jansen RT, Kuypers AW, Baadenhuijsen H, van den Besselaar AM, Cobbaert CM, Gratama JW, et al. Kalibratie 2000. Ned Tijdschr Klin Chem 2000;25:153–8.
7. 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;53:833–8.
8. Thelen MH, Jansen RT, Weykamp CW, Steigstra H, Meijer R, Cobbaert CM. Expressing analytical performance from multisample evaluation in laboratory EQA. Clin Chem Lab Med 2017;55:1509–16.
9. Thienpont LM, Van Uytfanghe K, De Grande LA, Reynders D, Das B, Faix JD, et al. On behalf of the IFCC Committee for Standardization of Thyroid Function Tests. Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. Clin Chem 2017;63:1248–60.
10. Klee GG. Clinical interpretation of reference intervals and reference limits. A plea for assay harmonization. Clin Chem Lab Med 2004;42:752–7.
11. Tate JR, Myers GL. Harmonization of clinical laboratory test results. eJIFCC 2016;27:005–14.
12. Leijnse B, Jacobs PH, Willebrands AF, Jansen AP. Eenheden in de klinische chemie en de blinde mol. Ned Tijdschr Geneesk 1972;116:114–5.
13. Kallner A, McQueen M, Heuck C. Consensus agreement conference on strategies to set global quality specifications in laboratory medicine. Stockholm. April 24–26, 1999. Scand J Clin Lab Invest 1999;59:475–6.
14. 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.
15. Clinical and Laboratory Standards Institute (CLSI). Characterization and qualification of commutable reference materials for laboratory medicine; Approved Guideline. CLSI document EP30-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2010.
16. Jansen R, Jassam N, Thomas A, Perich C, Fernandez-Calle P, Faria AP, et al. A category 1 EQA scheme for comparison of laboratory performance and method performance: an international pilot study in the framework of the Calibration 2000 project. Clin Chim Acta 2014;432:90–8.
17. Miller WG, Myers GL, Gantzer ML, Kahn SE, Schönbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
18. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
19. Braga F, Panteghini M. Verification of in vitro medical diagnostics (IVD) metrological traceability: responsibilities and strategies. Clin Chim Acta 2014;432:55–61.
20. Delanghe JR, Cobbaert C, Harmoinen A, Jansen R, Laitinen P, Panteghini M. Focusing on the clinical impact of standardization of creatinine measurements: a report by the EFCC Working Group on Creatinine Standardization. Clin Chem Lab Med 2011;49:977–82.
21. Drion I, Cobbaert C, Groenier KH, Weykamp C, Bilo HJ, Wetzels JF, et al. Clinical evaluation of analytical variations in serum creatinine measurements: why laboratories should abandon Jaffe techniques. BMC Nephrol 2012;13:133.
22. Jassam N, Yundt-Pacheco J, Jansen R, Thomas A, Barth JH. Can current analytical quality performance of UK clinical laboratories support evidence-based guidelines for diabetes and ischaemic heart disease? – A pilot study and a proposal. Clin Chem Lab Med 2013;51:1579–84.
23. Cobbaert C. Time for a holistic approach and standardization education in laboratory medicine. Clin Chem Lab Med 2017;55:311–3.
24. Müller MM. Implementation of reference systems in laboratory medicine. Clin Chem 2000;46:1907–9.
25. Müller MM. Traceability in laboratory medicine. Accred Qual Assur 2003;8:340–5.
26. Panteghini M. Traceability, reference systems and result comparability. Clin Biochem Rev 2007;28:97–104.
27. Dati F, Panteghini M, Apple FS, Christenson RH, Mair J, Wu AH. Proposals from IFCC Committee on Standardization of Markers of Cardiac Damage (CSMCD): strategies and concepts on standardization of cardiac marker assays. Scand J Clin Lab Invest Suppl 1999;230:113–23.
28. Ceriotti F. Harmonization initiatives in Europe. eJIFCC 2016;27:23–9.
29. Bais R, Armbruster D, Jansen RT, Klee G, Panteghini M, Passarelli J, et al. IFCC Working Group on Allowable Error for Traceable Results (WG-AETR). Defining acceptable limits for the metrological traceability of specific measurands. Clin Chem Lab Med 2013;51:973–9.
30. Jansen RT, Jansen AP. Standards versus standardised methods in enzyme assay. Ann Clin Biochem 1983;20:52–9.
31. Boerma GJ, Jansen AP, Jansen RT, Leijnse B, van Strik R. Minimizing interlaboratory variation in routine assays of serum cholesterol through the use of serum calibrators. Clin Chem 1986;32:943–947.

32. Jansen RT, Bullock DG, Vassault A, Baadenhuijsen H, De Leenheer A, Dumont G, et al. Between-country comparability of clinical chemistry results: an international quality assessment survey of 17 analytes in six European countries through existing national schemes. Ann Clin Biochem 1993;30:304–14.
33. Jansen R, Schumann G, Baadenhuijsen H, Franck P, Franzini C, Kruse R, et al. 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. Clin Chim Acta 2006;368:160–7.
34. Ross HA, on behalf of the Endocrinology Section and Project Group "Calibration 2000" of the SKML (Dutch Foundation for Quality Assessment in Clinical Laboratories). Reporting growth hormone assay results in terms of one consensus recombinant standard preparation offers less than optimal reduction of between method variation. Clin Chem Lab Med 2008;46:1334–5.
35. van Zelst BD, de Beer RJ, Neele M, Kos S, Kema IP, Tegelaers FP, et al. A multicenter comparison of whole blood vitaminB6 assays. Clin Chem Lab Med 2016;54: 609–16.
36. van den Besselaar AM, Haas FJ, van der Graaf F, Kuypers AW. Harmonization of fibrinogen assay results: study within the framework of the Dutch project 'Calibration 2000'. Int J Lab Hematol 2009;3:513–20.
37. Robijns K, Boone NW, Kuypers AW, Jansen RT, Neef C, Touw DJ. A multilaboratory commutability evaluation of proficiency testing material for carbamazepine and valproic acid: A study within the framework of the Dutch Calibration 2000 Project. Ther Drug Monit 2015;37:445–50.
38. Kluin-Nelemans JC, Van Wering ER, Van der Schoot CE, Adriaansen HJ, Van't Veer MB, Van Dongen JJ, et al. On behalf of the Dutch Cooperative Study Group on Immunophenotyping of Haematological Malignancies (SIHON). SIHONSCORE: a scoring system for external quality control of leukaemia/lymphoma immunophenotyping measuring all analytic phases of laboratory performance. Br J Haematol 2001;112:337–43.
39. van Toorenenbergen AW. Between-laboratory quality control of automated analysis of IgG antibodies against Aspergillus fumigatus. Diagn Microbiol Infect Dis 2012;74:278–81.
40. Klasen IS, Lentjes EG, Jol-van der Zijde CM, Backer ET, Kuypers AW, Baadenhuijsen H. The Calibration 2000 project: towards harmonisation of serum proteins using tertiary calibrators deduced from CRM470 (RPPHS). Ned Tijdschr Klin Chem 2000;25:159–62.
41. van den Besselaar AM, Haas FJ, van der Heij-Koene AJ. Kalibratie 2000: een tweelingstudie van kandidaat kalibratoren voor bepaling van fibrinogeen, factor VIII en antitrombine. Ned Tijdschr Klin Chem 2001;26:110.
42. 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.
43. 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.
44. Cobbaert C, Weykamp C, Franck P, De Jonge R, Kuypers A, Steigstra H, et al. Systematic monitoring of standardization and harmonization status with commutable EQA-samples-five year experience from the Netherlands. Clin Chim Acta 2012;414:234–40.
45. Weykamp C, Secchiero S, Plebani M, Thelen M, Cobbaert C, Thomas A, et al. Analytical performance of 17 general chemistry analytes across countries and across manufacturers in the INPUtS project of EQA organizers in Italy, the Netherlands, Portugal, United Kingdom and Spain. Clin Chem Lab Med 2017;55:203–11.
46. Ross HA, Lentjes EW, Menheere PM. The consensus statement on the standardization and evaluation of growth hormone and insulin-like growth factor assays lacks a recommendation to attempt efficacious harmonization. Clin Chem 2011;57:1463–4.
47. van den Besselaar AM, Haas FJ, Kuypers AW. Harmonisation of factor VIII: C assay results: study within the framework of the Dutch project 'Calibration 2000'. Br J Haematol 2006;132:75–9.
48. Van den Besselaar AM, Haas FJ, Kuypers AW. Harmonization and external quality assessment of antithrombin activity assays. Thromb Res 2012;129:187–91.
49. Robijns K, Boone NW, Jansen RT, Kuypers AW, Neef C, Touw DJ. Commutability of proficiency testing material containing tobramycin: a study within the framework of the Dutch Calibration 2.000 project. Clin Chem Lab Med 2017;55:212–7.
50. van der Burgt YE, Cobbaert CM, Dalebout H, Smit N, Deelder AM. Temperature-dependent instability of the cTnI subunit in NIST SRM2921 characterized by tryptic peptide mapping. J Chromatogr B Analyt Technol Biomed Life Sci 2012;902:147–50.
51. Rugaak LR, Romijn FP, Smit NP, van der Laarse A, Haas FJ, Meijer P, et al. Development of a mass spectrometry based method for targeted quantitation of clinically relevant proteoforms of antithrombin. Ned Tijdschr Klin Chem 2017;42:81.
52. Beastal GH, Brouwer N, Quiroga S, Myers GL. Traceability in laboratory medicine: a global driver for accurate results for patient care. Clin Chem Lab Med 2017;55:1100–11.
53. Armbruster D. Metrological traceability of assays and comparability of patient test results. Clin Lab Med 2017;37:119–35.
54. Baadenhuijsen H, Scholten R, Willems HL, Weykamp CW, Jansen RT. A model for harmonization of routine clinical chemistry results between clinical laboratories. Ann Clin Biochem 2000;37:330–7.
55. Baadenhuijsen H, Jansen RT, Weykamp CW, Steigstra H, Kuypers AW. A nationwide split patient and control sample experiment to check the commutability of control samples [Abstract]. Clin Chem Lab Med 1999;37:S279 (abstract T250).
56. Perich C, Ricós C, Alvarez V, Biosca C, Boned B, Cava F, et al. External quality assurance programs as a tool for verifying standardization of measurement procedures: Pilot collaboration in Europe. Clin Chim Acta 2014;432:82–9.
57. Weykamp C. HbA1c: a review of analytical and clinical aspects. Ann Lab Med 2013;33:393–400.
58. Cobbaert C, Baadenhuijsen H, Weykamp CW. Prime time for enzymatic creatinine methods in pediatrics. Clin Chem 2009;55:549–58.
59. Langlois MR, Descamps OS, van der Laarse A, Weykamp C, Baum H, Pulkki K, et al. EAS-EFLM collaborative project clinical impact of direct HDLc and LDLc method bias in hypertriglyceridemia. A simulation study of the EAS-EFLM Collaborative Project Group. Atherosclerosis 2014;233:83–90.

60. Weykamp C, Kuypers A, Bakkeren D, Franck P, van Loon D, Klein Gunnewiek J, et al. Creatinine, Jaffe, and glucose- another inconvenient truth. Clin Chem Lab Med 2015;53:e347–9.
61. Smit NP, Romijn FP, van den Broek I, Drijfhout JW, Haex M, van der Laarse A, et al. Metrological traceability in mass spectrometry-based targeted protein quantitation: a proof-of-principle study for serum apolipoproteins A-I and B100. J Proteomics 2014;109:143–61.
62. Ross HA, Lentjes EW, Menheere PM, Sweep CG. Endocrinology Section and Project Group "Calibration 2000" of the SKML (Dutch Foundation for Quality Assessment in Clinical Laboratories). Harmonization of growth hormone measurement results: the empirical approach. Clin Chim Acta 2014;432:72–6.
63. Schellenberg F, Wielders J, Anton R, Bianchi V, Deenmamode J, Weykamp C, et al. IFCC approved HPLC reference measurement procedure for the alcohol consumption biomarker carbohydrate-deficient transferrin (CDT): its validation and use. Clin Chim Acta 2017;465:91–100.
64. Weykamp C, Wielders JP, Helander A, Anton R, Bianchi V, Jeppsson JO, et al. Toward standardization of carbohydrate-deficient transferrin (CDT) measurements: III. Performance of serum native and serum spiked with disialotransferrin proves that harmonization of CDT assays is possible. Clin Chem Lab Med 2013;51:991–6.
65. Weykamp C, Wielders JP, Helander A, Anton R, Bianchi V, Jeppsson JO, et al. Harmonization of measurement results of the alcohol biomarker carbohydrate-deficient transferrin by use of the toolbox of technical procedures of the international consortium for harmonization of clinical laboratory results. Clin Chem 2014;60:945–53.
66. Klasen IS, Kuypers A, Weykamp C, Lentjes E, Jol-van der Zijde CM, Backer E, et al. Calibration 2000: results of a twin study for ten serum proteins. Ned Tijdschr Klin Chem 2001;26:68.
67. Kroot JJ, Van Herwaarden AE, Tjalsma H, Jansen RT, Hendriks JC, Swinkels DW. Second round robin for plasma hepcidin methods: first steps towards harmonisation. Am J Hematol 2012;87:1–7.
68. Cobbaert C, Michielsen E, Weykamp CW, Baadenhuijsen H, Van Dieijen-Visser M. Do NIST SRM 2921 and recombinant cTnI-based serum pools have potential to harmonize cTnI results? Ned Tijdschr Klin Chem Labgeneesk 2007;32:175–8.
69. van den Besselaar AM, van Rijn CJ, Cobbaert CM, Reiniers GL, Hollestelle MJ, Niessen RW, et al. Fibrinogen determination according to Clauss: commutability assessment of international and commercial standards and quality control samples. Clin Chem Lab Med 2017;55:1761–9.
70. de Metz M, van den Berg GA, van Duijnhoven JL, Berends F, Verhoef NJ, Kuijpers AW. Calibration 2000 project. Hemocytometry. Ned Tijdschr Klin Chem 2001;26:78.
71. Gratama JW, Kraan J, Keeney M, Granger V, Barnett D. Reduction of variation in T-cell subset enumeration between 55 laboratories using single-platform, 3- or 4-color flow cytometry based on CD45 and SSC based gating of lymphocytes. Cytometry B Clin Cytom 2002;50:92–101.
72. Gratama JW, Kraan J, Keeney M, Sutherland DR, Granger V, Barnett D. Validation of the single-platform ISHAGE method for CD34+ hematopoietic stem and progenitor cell enumeration in an international multicenter study. Cytotherapy 2003;5:55–65.
73. Levering WH, Wind H, Sintnicolaas K, Hooijkaas H, Gratama JW. Flow cytometric HLA-B27 screening: cross-reactivity patterns of commercially available anti-HLA-B27 monoclonal antibodies with other HLA-B antigens. Cytometry B Clin Cytom 2003;54B:28–38.
74. Levering WH, Wind H, Hooijkaas H, Sintnicolaas K, Brando B, Gratama JW. Flow cytometric screening for HLA-B27 on peripheral blood lymphocytes. J Biol Regul Homeost Agents 2003;17:241–6.
75. Raven S, Hautvast J, van Steenbergen J, Akkermans R, Weykamp C, Smits F, et al. Diagnostic performance of serological assays for anti-HBs testing: results from a quality assessment program. J Clin Virol 2017;87:17–22.
76. IFCC Working group Apolipoproteins by Mass Spectrometry (WG-APO MS). http://www.ifcc.org/ifcc-scientific-division/sd-working-groups/wg-apo-ms/.
77. Van Duijn MM, Jacobs JF, Wevers RA, Engelke UF, Joosten I, Luider TM. Quantitative measurement of immunoglobulins and free light chains using mass spectrometry. Anal Chem 2015;87:8268–74.
78. Brouwer N, Den Elzen W, Thelen M, Haagen I, Cobbaert C, Number: national reference intervals and decision limits in The Netherlands using a 'big data' approach. Clin Chem Lab Med 2017;55(Special Suppl.):S1059.