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

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 true-ness 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.

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).

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 inter- and intralaboratory differences, especially differences in analytical set-points, which 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 inter- and intralaboratory differences, especially differences in analytical set-points, which can markedly affect the clinical interpretation of tests [10].

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 HbA1c 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 focussed 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.

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

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

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...
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].

HbA1c is another test that needed harmonization [57]. Also for HbA1c important results were achieved. An international HbA1c network has been established for global standardization of HbA1c. In the Netherlands, two laboratories operate the international reference method for HbA1c 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 HbA1c 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
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 HbA1c (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-I). The diagnosis growth hormone deficiency largely depends on the hGH test result. Because of the relevance of growth hormone treatment in children with short stature, 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)/).

**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, α1-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 focusses on method robustness and harmonization. For example, first steps are underway for hepcidin [67].

The use of NIST SRM 2921 and recombinant cTnl-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
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.
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


