FACULTY OF
PHARMACEUTICAL SCIENCES

Modern tools for improving comparability and stability of in vitro diagnostic tests

Kenneth Goossens

Promotor(s):

Prof. Emer. Dr. Linda Thienpont

Prof. Dr. Tom Coenye

Thesis submitted to obtain the degree of Doctor in Pharmaceutical Sciences

Members of the Examination Committee

Chairman: Prof. Dr. Apr. Christophe Stove

Laboratory of Toxicology, Department of Bioanalysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

Secretary: Prof. Dr. Thomas De Beer

Laboratory of Pharmaceutical Process Analytical Technology, Department of Pharmaceutical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

Dr. Klin. Biol. Peter Declercq

Clinical Laboratory , vzw Jessa ziekenhuis, Hasselt, Belgium

Apr. Klin. Biol. Timothy Ghys

Clinical Laboratory, AZ Sint-Lucas, Ghent, Belgium

Prof. Dr. Sverre Sandberg

The Norwegian Quality Improvement of Primary Care Laboratories (Noklus), Haralds-plass Deaconess Hospital, Bergen, Norway;

Department of Global Health and Primary Care, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway;

Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway.

Prof. Dr. Veronique Stove

Clinical Laboratory, University Hospital of Ghent, Ghent, Belgium

Table of contents

List of abbreviations

AACC	American Association for Clinical Chemistry
ALP	Alkaline Phosphatase
ALT	Alanine Aminotransferase
AMTM	All Manufacturer Trimmed Mean
AST	Aspartate Aminotransferase
CLSI	Clinical & Laboratory Standards Institute
C-STFT	Committee for Standardization of Thyroid Function Tests
CV	Coefficient of Variation
DGKC	Deutsche Gesellschaft für Klinische Chemie
EFLM	European Federation of Clinical Chemistry and Laboratory Medicine
EQA	External Quality Assessment/Assurance
FDA	Food and Drugs Administration
FT4	Free Thyroxine
FW	Forward
GGT	Gamma-Glutamyltransferase
GUM	Guide to the expression of Uncertainty in Measurement
IFCC	International Federation for Clinical Chemistry and Laboratory Medicine
ISE	Ion Selective Electrode
ISO	International Organization for Standardization
IQC	Internal Quality Control
JCTLM	Joint Committee for Traceability in Laboratory Medicine
LDH	Lactate Dehydrogenase
LIS	Laboratory Information System
MC	Master Comparison
MM	Moving Median
MU	Measurement Uncertainty
NCEP	National Cholesterol Education Program
PP	Pyridoxal Phosphate
QC	Quality Control
QI	Quality Indicator
RE	Random Error
RMP	Reference Measurement Procedure
RMP-c AMTM	Reference Measurement Procedure-Corrected All Manufacturer Trimmed Mean
RV	Reverse
SD	Standard Deviation
SE	Systematic Error
SI	Système International d’ Unités
SRE	Sample-Related Effect
TE	Total Error
TSH	Thyroid Stimulating Hormone
VIM	Vocabulaire International de Métrologie/ International Vocabulary of Metrology

List of definitions

Coefficient of variation (CV):

A measure of dispersion defined as the ratio of the standard deviation to the mean (often expressed as a percentage).

(Daily) patient medians:

Median of results of patient samples, preferably outpatient samples, which are measured on a single day in a certain laboratory on a certain instrument.

Hypo- and hyper-flagging rates:

The daily number of results, expressed as percentage relative to the total number, that are automatically flagged when they are either lower (hypo) or higher (hyper) than the locally used decision limits or reference interval.

Long-term median:

A statistical measure that represents the median calculated from all daily patient medians (Percentiler) or daily flagging rates (Flagger) provided by the laboratory to which the graph applies. In the Percentiler and Flagger graphs this is indicated with the long-broken horizontal grey line.

Moving average:

A statistical measure which is commonly used with time series data in order to create charts that show whether a significant trend is visible by flattening out noise. To do so a series of averages is calculated of different subsets of the full dataset.

Moving median:

This measure is similar to the moving average but instead of a series of averages, a series of medians is calculated, which is less sensitive towards outliers. In the Percentiler these moving medians are calculated for daily patient medians, in the Flagger for daily hypo-and hyper-flagging rates (each time grouped per 5, 8 or 16 days).

Robust CV:

The robust CV is calculated as 100 * (median absolute difference x 1.4826)/(median). The robust CV is less sensitive towards outliers, because it uses the median absolute difference as measure of the variability of quantitative data. This median absolute difference is calculated by taking the median of the absolute differences of each measurement in the dataset with the median of that dataset.

Sy/x (standard error of estimate):

A statistical measure for the accuracy of predictions made with a regression line. The value either represents the within-run imprecision (when a laboratory is compared with its peer) or the combined imprecision and sample-related effects (when a laboratory is compared with a trueness based target).

Introduction

Highlights

In healthcare, just like in any other discipline, a measurement is not an absolute concept, but it delivers an estimate of the “true” value. This is due to the fact that a measurement is always accompanied by a certain amount of error, which consists of a one-sided systematic error and a two-sided random error. These two types of error, which by definition have different origins, make up for the total error of the measurement. Making a measurement error on an *in vitro* diagnostic (IVD) instrument is unfortunately inevitable. Therefore, in order to avoid inaccurate patient management, the total error should always be minimized and not exceed a certain quality specification. A clinical laboratory can apply internal quality control (IQC) and external quality assessment (EQA) tools to control the analytical quality it manages when it measures patient samples.

This chapter describes:

- The concept “error”: the different error components and error models.
- Quality specifications – what is a significant error?
- Established quality management tools in the clinical laboratory: IQC and EQA

1. Analytical Quality (Assessment) – Measurement introduces error

Introducing significant (analytical) errors in a clinical laboratory environment can have a harmful effect on patient care and health economics. These errors can, for example, lead to falsely interpreting analytical runs and misclassification of patient results (1). A report from the Institute of Medicine from the United States (US) indicates that medical errors are costing the US healthcare system $17 billion each year (2). Although In Vitro Diagnostics (IVD) only contribute to a small part of these errors, performing measurements on an IVD instrument will always introduce a certain amount of error, that could lead to results which are not necessarily stable or comparable over time and location. Table 1 shows an educational example of a patient who is being monitored for his/her alanine aminotransferase (ALT) values on different platforms.

Table 1 Educational example of ALT measurements (fictional case).

Date	Measuring system	ALT (U/L)	Patient/Physician communication
Jan 2010	Cobas S1	20	Patient monitors his/her ALT out of curiosity, even though the values are considered healthy
April 2010	Cobas S1	25	Patient asks about difference, physician explains biological variation
July 2010	Cobas S1	26	Patient is happy, values are normal and stable
Oct 2010	Cobas S2	30	Patient gets nervous, physician says short-term elevations are normal; patient requires retest
Oct 2010	Cobas S2 (new lot)	35	Patient stays nervous, physician repeats that short-term elevations are normal
Jan 2011	Cobas S2	34	Patient is happy, values are normal and stable
July 2011	Architect	45	Patient is near crisis, new physician explains report and new reference intervals; patient requires retest
July 2011	Architect	46	Patient satisfied about health, but dissatisfied about physician, extends retest period
Jan 2012	Architect	44	Patient calmed down
Oct 2012	Vitros	60	Patient in “panic” (new town, new physician); physician repeats above stories; patient requests retest
Oct 2012	Vitros	59	Patient dissatisfied, but accepts
Oct 2013	AU	31	Patient starts to amuse him/herself
Oct 2014	Cobas	27	Patient becomes interested in the topic
Oct 2014	Patient stops	None	Patient writes a thesis

The patient had been tested before for ALT by his general practitioner (GP) because of certain clinical symptoms. Although testing was negative, the patient was so concerned about his health status that he urged his GP to continue the monitoring of his ALT values. Over time the measured ALT value changes significantly due to a combination of biological variation and a number of quality-related problems. The Westgard biodatabase (3) indicates that the typical within-subject biological variation for ALT equals 19.40%. Assuming the numbers from that source are correct, the differences shown in Table 1 can't simply be explained by biological variation. And, although the mentioned values in Table 1 are fictional, they still represent realistic numbers, which is reflected in Figure 1. This dotplot shows the outcome when a single sample is measured for ALT in 126 different laboratories (data from our Master Comparison survey). It shows huge measurement variation for ALT and indicates that a twofold difference between laboratories is not impossible. Note, here abstraction is made from the aforementioned biological variation.

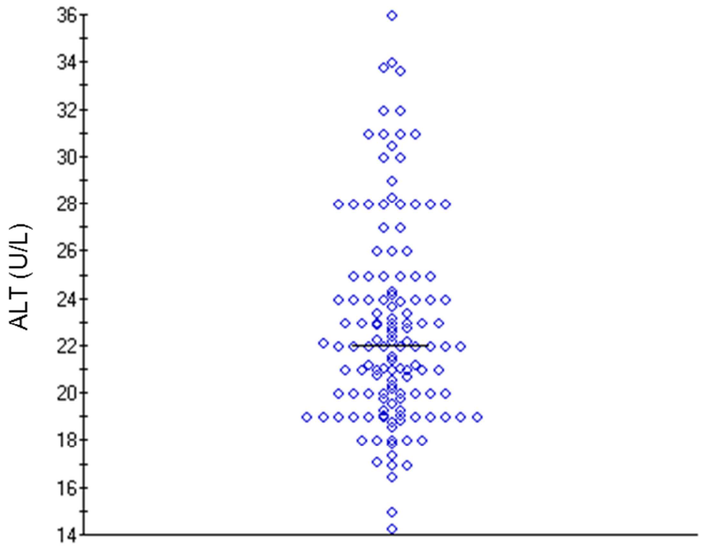


Figure 1. Dotplot showing the distribution of measurements for a single-donation sample (126 laboratories using different IVD systems). Measurement values extend from 14 to 36 U/L, with an average value of 22 U/L.

This dotplot is actually a nice demonstration which indicates that a measurement is not an absolute concept – as some patients may believe – but an estimate of the true value. The correctness of that estimate depends on the size of the error which is made during the patient management process. Errors can occur at three different stages: the pre-analytical, analytical and post-analytical phase (4). The pre- and post-analytical phases consist of every aspect in patient management besides the clinical laboratory. In short, the pre-analytical phase includes the collection of patient samples and its transportation, whilst the post-analytical phase includes the interpretation and utilization of laboratory information (5). This thesis, however, will mostly focus on the analytical phase, which consists of the core laboratory work under control of the laboratory staff. Errors introduced in this phase relate to the reliability of the applied instrumentation.

The different error components

As already mentioned, a measurement is always accompanied by a certain amount of error. The total error (TE) can be described as a combination of random error (RE) or (im)precision, systematic error (SE) or bias, and in some cases also interferences (see Figure 2). The International Vocabulary of Metrology (VIM) defines measurement accuracy (which relates to the TE) as: “The closeness of agreement between a measured quantity value and a true quantity value of a measurand”. Measurement accuracy consists of both measurement trueness (which relates to the SE) and measurement precision (which relates to the RE). Trueness is defined as: “closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value”, and precision as: “closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions” (6).

The total imprecision or RE consists of a combination of imprecision and sample-related effects (SRE’s). The imprecision is a measure for the degree of variation of a result in a set of replicate measurements (7). It is an inevitable characteristic of every measurement, where results lie on both sides of a mean value. It is usually expressed as a standard deviation (SD) or coefficient of variation (CV). One usually makes a distinction between within-run imprecision and between-run imprecision of a method, which can quite easily be assessed with available guidelines such as the EP05-A3 protocol (8,9). When the RE of a method is estimated with the Sy/x value after linear regression analysis, SRE’s can increase the RE due to the inherent variation of the measured samples (Sy/x = the standard deviation of the residuals). This additional RE or lack of fit error causes the data to not strictly fit the estimated regression line (10). In contrast to the measurement imprecision, retesting and increasing the number of replicates will not significantly lower the error caused by SRE’s.

A bias or SE is defined as: “the mean that would result from an infinite number of measurements carried out under repeatability conditions minus a true value of the measurand” (7). As shown in Figure 2, this bias can be introduced at several levels of the measurement process (manufacturer, laboratory, module of the instrument, lot, and calibration). Again, several guidelines (e.g. the EP09-A3 protocol) are available to assess the size of the bias (11). Interference, finally, is defined as: “a SE in the measure of signal caused by the presence of concomitants in a sample” (7,12). The interference can be assessed with the EP07-A3 protocol (13).

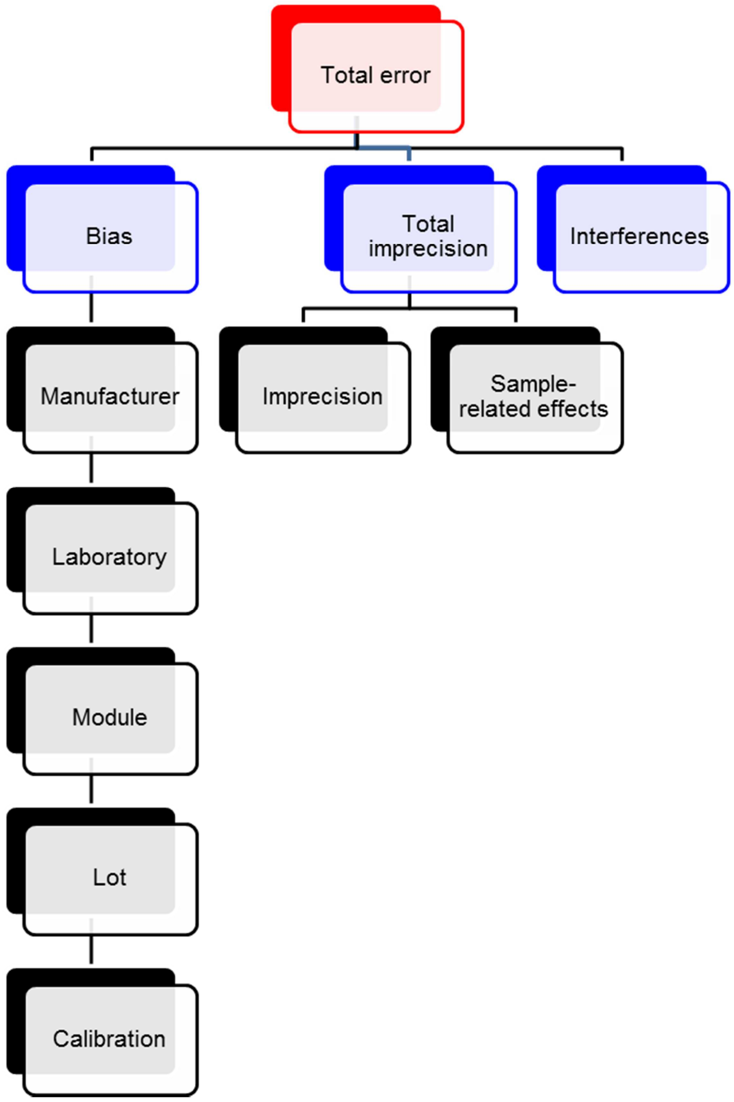


Figure 2. Overview of total error components influencing the test result.

Trueness – Establishing a traceability chain

In the hypothetical case that a SE is non-existent, the mean that would result from an infinite number of measurements carried out under repeatability conditions equals the "true value". In order to realize the trueness of a routine measurement method or IVD instrument, one needs to establish its traceability to a higher order "reference" using appropriate materials (14,15). The International Organization for Standardization (ISO) introduced the traceability concept as a means to link the result of patient samples to a commonly accepted reference, in order to make them comparable over time and location, and across systems (16). In the ideal case, all routine methods are traceable – also called standardized – to a trueness-based higher order primary reference material and/or reference measurement procedure (RMP), both realizing the Système International d' Unités (SI).

The traceability chain itself starts with the definition of the measurand and the unit of expression for measurement results (preferably SI), and consist of an unbroken chain of calibrations (depicted in Figure 3). Each material from this chain (blue frames in Figure 3) is used to calibrate a lower-order measurement procedure (red frames in Figure 3). The calibrated measurement procedure is, in turn, used to assign a value to the lower calibrator. Given that all elements from the chain are present, calibration starts with a primary reference material used by reference laboratories to calibrate the highest order RMP, and goes all the way down, via the instrument manufacturers, to the routine clinical laboratory. Meanwhile, the inevitable uncertainty of the measurement increases after each calibration step, to reach the highest uncertainty at the level of the clinical laboratory.

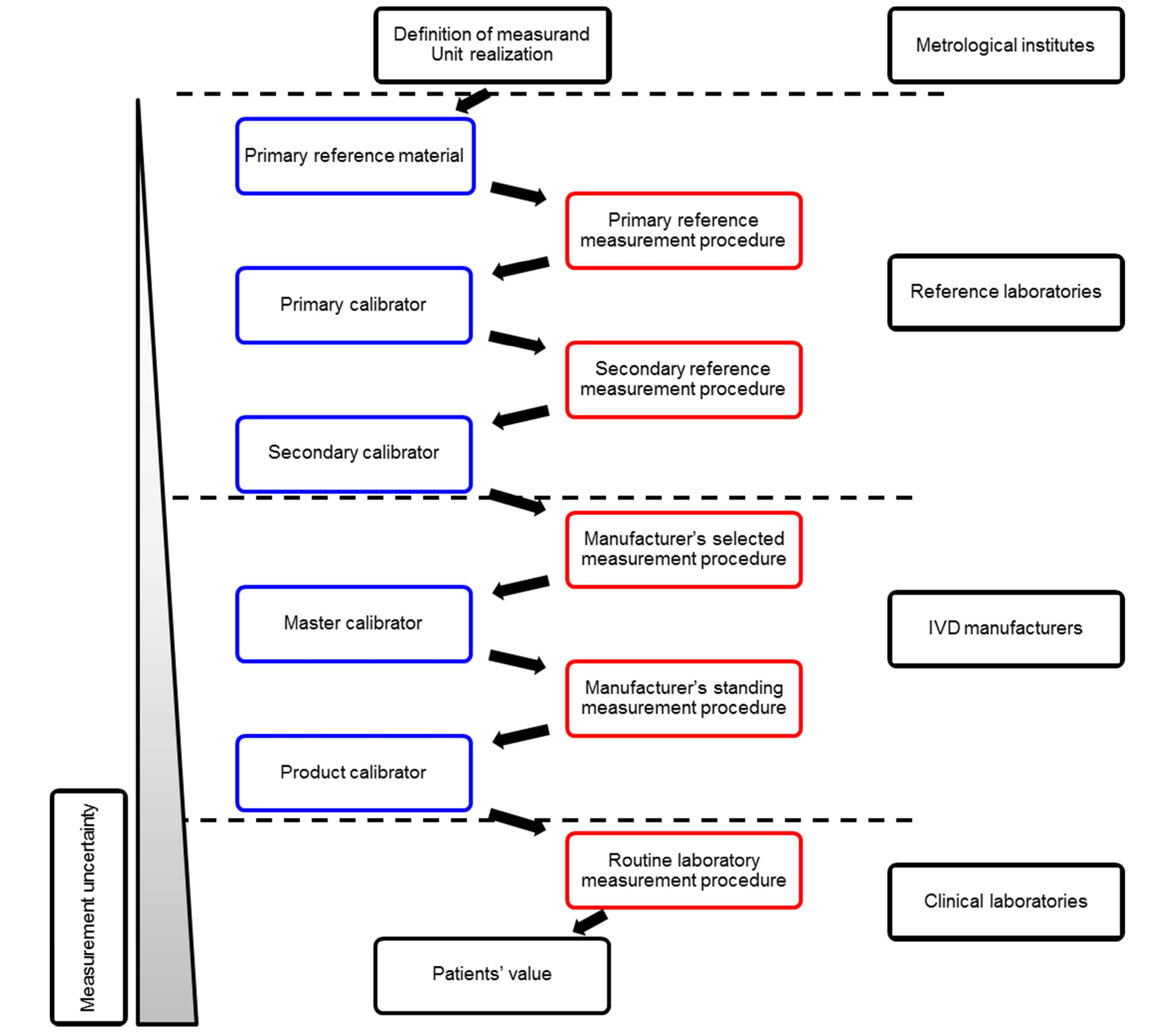

Definition of measurand
Unit realization
Metrological institutes
Primary reference material
Primary reference measurement procedure
Reference laboratories
Primary calibrator
Secondary reference measurement procedure
Secondary calibrator
Manufacturer's selected measurement procedure
Master calibrator
IVD manufacturers
Manufacturer's standing measurement procedure
Product calibrator
Measurement uncertainty
Routine laboratory measurement procedure
Clinical laboratories
Patients' value

Figure 3. Metrological traceability chain. The traceability chain consist of an unbroken chain of calibrations. Each material (blue frames) is used to calibrate a lower-order measurement procedure (red frames). The calibrated measurement procedure is, in turn, used to assign a value to the lower calibrator.

Two problems can arise related to the concept traceability. Either the chain is incomplete, or an artificial bias is introduced by one of the applied calibrating materials. To account for the former issue different concepts of the traceability chain have been established (see Vesper *et al*; 14). A returning problem is the unavailability of a higher-order reference material and/or method. This means that SI-traceability or standardization of routine measurements is not possible, but this doesn't mean that equivalent results across different methods cannot be obtained. The routine methods can still be harmonized to a lower-order reference material or to a surrogate target (17). One of this surrogate targets can, for example, be an "All Method/Manufacturer Trimmed Mean" or AMTM (18). Lack of both standardization and harmonization will lead to a lack of comparability between different methods available on the market.

A second problem arises when a bias is introduced due to a matrix effect. This problem is referred to as "non-commutability" (19,20). This is typically due to the fact that the measurand (or quantity which is intended to be measured) is contained in a different matrix for the calibrator/reference material compared to patient samples. When both of these samples (reference materials and patient samples) are measured with two or more analytical methods, the reference materials can only be considered as commutable when they have interassay properties comparable to the properties of the patient samples (21). Non-commutable reference materials should never be used for calibration purposes as they will hamper the traceability chain and jeopardize a trueness-based value.

Apart from establishing a traceability chain, one should also select measurement principles which are as insensitive as possible to interferences to be specific, and select stable, high quality instrumentation (i.e. without system drifts or shifts) in order to minimize the SE and guarantee trueness.

The different error models

As explained in the VIM vocabulary, RE and SE (or precision and trueness) are by definition different. They are combined in different ways to calculate the TE (22,23). The following models exist:

1. The variance model treats the two types of error as two similar entities and combines them by error propagation (24).

   $$\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{analytical}} + |\text{bias}|^2}$$

2. In the measurement uncertainty model (MU model) described by the Guide to the expression of Uncertainty in Measurement (GUM), the error around the estimated quantity is expressed as uncertainty. The GUM states that all known components of error should be evaluated and if possible corrected for. For example, known bias components should be eliminated by applying correction factors. The uncertainty of that correction is added to all other forms of uncertainty by error propagation. Unknown bias components are transformed into variances assuming a rectangular distribution with the same probability for any value of the presumed interval (25). This model is mostly used in the Quality Control (QC) environment of IVD manufacturers and reference laboratories.

3. Thirdly, the TE or linear model (also referred to as the Westgard model) combines bias and imprecision via a linear relationship (26). This model is mostly used in the QC environment of clinical laboratories.

   $\text{TE} = |\text{bias}| + z.\sigma_{\text{analytical}}$
   (with $z = 1.65$ for one-tailed probability or 1.96 for two-tailed probability)

4. Finally, the combined model includes the biological variation component in the equation and is, therefore, a useful model to be applied when inferring quality goals (discussed in part 2 of the introduction). This model is more complex, and discussing it is beyond the scope of this introduction. Further information can be found in the publication of Gowans *et al* (27).

The TE and MU model are most frequently used in a QC environment. Nonetheless, there is some debate on the correct use of these different types of models (28-31). Westgard promotes a clear separation between the TE and MU model. He proposes to use the TE model in the QC environment of a clinical laboratory, because it allows to use one error budget (indicated with TE limits or a total allowable error) which laboratories can spend for both bias and imprecision without having to differentiate between both type of errors. On the other hand he proposes to use the MU model for quality management by manufacturers, reference laboratories and metrological institutes to assure the traceability and comparability of results across clinical laboratories (28). Oosterhuis *et al* find this clear separation too simplistic. They propose to improve the TE model by implementing uncertainty component estimations to take into account all sources of variation when handling patient samples, i.e. not only analytical variation, but also variation within and between reagent lots and calibrations, within and between biological variation, pre-analytical variation, etc. The idea behind this concept is to include all sources of variation to get an idea on the total MU when comparing patient samples with either a previous result, with a population reference interval or with a clinical decision point. The MU model would, thereby, ease the identification of patient samples which are significantly different from another result for the same quantity (29,30). Currently, however, the TE model is most frequently used at the level of the clinical laboratory, because removing all known bias components – as suggested by the MU model – remains a cumbersome to impossible task for the clinical laboratory.

2. Quality specifications – What values are acceptable for bias and imprecision?

One problem which frequently arises in the investigation of laboratory error is how to define what actually constitutes significant error. Introduction of measurement error is unavoidable, therefore the question remains how large this error can be without compromising medical decisions? Or expressed in an alternative manner: what quality is sufficient for patient care? To answer this question, one has to rely on quality specifications which define a reasonable budget to make measurement errors (32). This leads to the next question: what values for quality specifications should be used? Inferring these values unequivocally is not that simple because different strategies have been proposed to calculate specifications or limits. A first landmark on how to deal with these different strategies was proposed at the 1999 Stockholm conference organized by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). This conference derived a Consensus Statement on analytical performance specifications, which has set a clear hierarchy on the strategies which can be applied to calculate the specifications (33,34). This hierarchy is addressed in Table 2 and shows five hierarchal levels.

Table 2. The hierarchy of strategies/models to set analytical quality specifications as decided on the 1999 Stockholm conference (33).

1. Evaluation of the effect of analytical performance on clinical outcomes in specific clinical situations.
2. Evaluation of the effect of analytical performance on clinical decisions in general.
    a. Data based on the components of biological variation.
    b. Data based on analysis of clinicians' opinions.
3. Published professional recommendations.
    a. From national and international expert bodies.
    b. From expert local groups or individuals.
4. Performance goals set by:
    a. Regulatory bodies
    b. Organizers of External Quality Assessment (EQA) schemes
5. Goals based on the current state of the art.
    a. As demonstrated by data from EQA or Proficiency Testing Schemes.
    b. As found in current publications on methodology

At the first strategic conference of the EFLM on “Defining analytical performance goals 15 years after the Stockholm conference on quality specifications in laboratory medicine” held in Milan on November 2014, the consensus agreement on quality specifications was revisited (35). This revision mostly led to a simplification of the hierarchy from five to three models:

1. The first model is based on the effect of clinical performance on clinical outcomes. These types of specifications are called outcome-based quality limits and they are calculated using either direct or indirect outcome studies. Direct outcome studies are extremely valuable because they reflect the clinical needs of patients, but they are also difficult to perform because clinical tests are only indirectly linked to the clinical outcome of that test (36). That explains why there are only a limited number of examples available on these types outcome studies (e.g. 37-40). Indirect outcome studies either rely on simulations to model the probability of clinical outcome (41), or on clinicians’ and/or experts’ opinion, e.g. by providing questionnaires to clinicians (42).
2. The second model is based on the biological variation of the measurand. This approach attempts to minimize the ratio of analytical to biological variation (43). Two of the most frequently used concepts for desirable performance specifications have been described by Harris (44) and Gowans (27). Harris *et al* discussed that, in order for the analytical variation to add maximally 12% to the total test variability, the analytical variation should be smaller than 50% of the within-subject biological variation ($CV_A \leq 0.5CV_I$). Gowans at her turn showed that, in order for laboratories with a homogeneous population to use the same reference interval, the bias should be lower than 25% of the error propagation of the within- and between biological variation ($\text{bias} \leq 0.25(CV_I^2 + CV_G^2)$). Other models based on biological variation exist, but discussing these is beyond the scope of this introduction.

   The advantage for models based on biological variation is that they are simple to use and they can be applied to most measurands for which within-subject and between-subject biological variation data have been established. There are, however, several limitations to this approach: (i) sometimes the biological variation is quite narrow leading to limits which are too stringent, (ii) when the biological variation is too broad, the limits will be too generous, (iii) different studies provide different outcomes on biological variation, and (iv), this model doesn't consider the necessity for different requirements in different concentration ranges.

3. The third and last model is based on state-of-the-art performance. These specifications show what is currently achievable at the highest level of analytical performance. These limits are usually assessed with empirical methods, for example by performing an EQA survey. Different studies, however, usually don't provide a consensus on the goal-setting.

   Depending on the applied empirical model/concept, these limits can either represent what is currently achievable by the best instrument and/or assay on the market or represent a certain quality which most laboratories can achieve. Laboratories which are not able to meet this level of quality are expected to verify and, if needed, change their practice. State-of-the-art performance limits can, in addition, be compared with the limits from other models to identify the technology which should be improved in order to meet the requirements to suffice for clinical needs.

Although in theory one should always choose the highest possible model in the hierarchy, certain models will be more suitable for certain measurands than for others. That's why there is currently still no clarity on which model to be used for which test. Therefore, a task force has been created by the EFLM to allocate different tests to different models (45).

3. How to address problems with analytical quality in the clinical laboratory

Typically, the individual patient's health status is monitored in a specific laboratory which uses a module of an IVD instrument with a particular assay. This means that a measurement result for the individual patient will comprise the full test variation or measurement error extending from the manufacturer over laboratories to system sites. This also means that a whole range of quality-related problems can occur upon measurement of a patient's sample in a clinical laboratory including interference errors, pre-analytical variation, bias-related issues, stability problems, lot- and calibration variation... The list of potential errors is quite extensive and measurement errors which are too large will lead to the inability to reliably examine a patient's health status or monitor the patient over time and across sites. Luckily, tools to map these problems are available in the quality monitoring discipline.

First of all, there is the ISO 15189 accreditation standard which specifies requirements for the competence and quality of clinical laboratories (46,47). The goal of this international standard for common use is to enhance the credibility of accredited clinical laboratories, and to provide a minimum standard of quality testing. To achieve this international standard, laboratory quality management typically relies on IQC and EQA (or proficiency testing) tools. These are not only used to identify (the cause of) erroneous results, but also provide a means to correct for errors (4). IQC in the clinical laboratory – just like any other analytical laboratory – involves a continuous evaluation of the laboratory's own analytical methods to monitor day-to-day precision and accuracy of a given assay. The theory behind IQC is that the laboratory runs QC samples in parallel with routine patient samples. The results for the QC samples are plotted in a control chart to verify whether the measurement procedure performs stable within given specifications. If the QC results fall outside the specifications, corrective action might be necessary (48). EQA, on the other hand tries to ensure interoperable/comparable results by comparing a laboratory to its peer, i.e. a group of laboratories using instruments with similar technology, or to a trueness-based reference system. These surveys are usually overseen by a third party and imply sending a (small) set of samples to all survey participants. These are expected to handle and measure the EQA samples in the same manner as the patient samples. The results of the measured samples are returned to the third party, which then processes the results from all survey participants and provides feedback on the quality of each laboratory (49).

References

1. Kinns H, Pitkin S, Housley D, Freedman DB. Internal quality control: best practice. J Clin Pathol 2013;66(12):1027-32.
2. Kohn LT, Corrigan JM, Donaldson MS, for the Institute of Medicine. To err is human: building a safer health system. National Academy Press 2000.
3. Westgard QC. Desirable specifications for total error, imprecision, and bias, derived from biologic variation. https://www.westgard. com/biodatabase1.htm (accessed January 2014)
4. Plebani M. The detection and prevention of errors in laboratory medicine. Ann Clin Biochem 2010;47(Pt 2):101-10.
5. Hawkins R. Managing the pre- and post-analytical phases of the total testing process. Ann Lab Med 2012;32:5-16.
6. JCGM 200:2012. International vocabulary of metrology – basic and general concepts and associated terms (VIM). Geneva: International Organization for Standardization; 2007. http://www.bipm.org/en/publications/guides/vim.html (Accessed June 2016).
7. International Union of Pure and Applied Chemistry (IUPAC). Compendium of Chemical Terminology Gold Book. 2014. http://goldbook.iupac.org/ (accessed June 2016).
8. Chesher D. Evaluating Assay Precision. Clin Biochem Rev 2008;29(Suppl 1):S23–S26.
9. Wayne, PA: Clinical and Laboratory Standards Institute (CLSI). Evaluation of precision of quantitative measurement procedures; approved guideline - third edition. CLSI document EP05-A3. 2014.
10. Long T. Statistical power in the detection of matrix effects. Arch Pathol Lab Med 1993;117:387-92.
11. Wayne, PA.: Clinical and Laboratory Standards Institute (CLSI). Measurement procedure comparison and bias estimation using patient samples; approved guideline - third edition. CLSI document EP09-A3. 2013.
12. Kroll MH, Elin RJ. Interference with clinical laboratory analyses. Clin Chem 1994;40:1996-2005.
13. Wayne, PA: Clinical and Laboratory Standards Institute (CLSI). Interference testing in clinical chemistry; approved guideline - second edition. CLSI document EP07-A2. 2005.
14. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55(6):1067-75.

15. Thienpont LM. Accuracy in clinical chemistry - who will kiss Sleeping Beauty awake? Clin Chem Lab Med 2008;46(9):1220-22.

16. Geneva: International Organization for Standardization (ISO). In vitro diagnostic medical devices – measurement of quantities in biological samples – metrological traceability of values assigned to calibrators and control materials. ISO 17511:2003.

17. Miller W, Myers GL, Lou Gantzer M, Kahn SE, Schönbrunner ER, Thienpont LM, Bunk DM, Christenson RH, Eckfeldt JH, Lo SF, Nübling CM, Sturgeon CM. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57(8):1108-17.

18. Thienpont LM, Van Houcke SK. Traceability to a common standard for protein measurements by immunoassay for in-vitro diagnostic purposes. Clin Chim Acta 2010;411(23-24):2058-61.

19. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52(4):553-4.

20. Miller WG, Myers GL. Commutability still matters. Clin Chem 2013;59(9):1291-3.

21. Wayne, PA. CLSI. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline. CLSI document EP30-A. Clinical and Laboratory Standards Institute; 2010.

22. Petersen PH, Stöckl D, Westgard JO, Sandberg S, Linnet K, Thienpont L. Models for Combining Random and Systematic Errors. Assumptions and Consequences for different Models. Clin Chem Lab Med 2001;39(7):589-95.

23. Hyltoft Petersen P, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using Guideline-Driven Medical Decision Limits. Clin Chim Acta 2014;430:1-8.

24. Harris EK. Proposed goals for analytical precision and accuracy in single-point diagnostic testing. Theoretical basis and comparison with data from College of American Pathologists proficiency surveys. Arch Pathol Lab Med 1988;112(4):416-20.

25. BIPM, IEC, IFCC. Evaluation of measurement data – guide to the expression of uncertainty in measurement GUM. JCGM 1002nd ed. 2008.

26. Westgard JO, Barry PL. Cost-effective quality control: managing the quality and productivity of analytical processes. AACC Press 1986.

27. Gowans EM, Hyltoft Petersen P, Blaabjerg O, Horder M. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 1988;48:757-64.

28. Westgard JO. Useful measures and models for analytical quality management in medical laboratories. Clin Chem Lab Med. 2016;54(2):223-33.

29. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? Clin Chem Lab Med. 2016;54(2):235-9.
30. Jones GRD. Measurement uncertainty for clinical laboratories – a revision of the concept. Clin Chem Lab Med. 2016;54(8):1303-7.
31. Panteghini M, Sandberg S. Total error vs. measurement uncertainty: the match continues. Clin Chem Lab Med. 2016;54(2):195-6.
32. Krouwer JS. Setting performance goals and evaluating total analytical error for diagnostic assays. Clin Chem 2002;48(6 Pt 1):919-27.
33. Kenny D, Fraser CG, Petersen P, Kallner A. Consensus agreement. Scand J Clin Lab Invest 1999;59:585.
34. Fraser CG. The 1999 Stockholm Consensus Conference on quality specifications in laboratory medicine. Clin Chem Lab Med 2015;53(6):837-40.
35. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, Petersen PH, Schimmel H, Sikaris K, Panteghini M. 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(6):833-5.
36. Horvath AR, Bossuyt PM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WD, Lennartz L, Cobbaert CM, Ebert C, Lord SJ; Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine. Setting analytical performance specifications based on outcome studies - is it possible? Clin Chem Lab Med 2015;53(6):841-8.
37. Larsen ML, Fraser CG, Petersen PH. A comparison of analytical goals for haemoglobin A1c assays derived using different strategies. Ann Clin Biochem 1991;28(Pt 3):272-8.
38. Jenny RW. Analytical goals for determinations of theophylline concentration in serum. Clin Chem 1991;37(2):154-8.
39. Petersen PH, Hørder M. Ways of assessing quality goals for diagnostic tests in clinical situations. Arch Pathol Lab Med 1988;112(4):435-43.
40. Petersen PH, Hørder M. Influence of Analytical Quality on Test Results. Scand J Clin Lab Invest 1992;52 Suppl 208:65-87.
41. Petersen PH. Performance criteria based on true and false classification and clinical outcomes. Influence of analytical performance on diagnostic outcome using a single clinical component. Clin Chem Lab Med 2015;53(6):849-55.
42. Thue G, Sandberg S. Analytical performance specifications based on how clinicians use laboratory tests. Experiences from a post-analytical external quality assessment programme. Clin Chem Lab Med 2015;53(6):857-62.

43. Fraser CG. General strategies to set quality specifications for reliability performance characteristics. Scand J Clin Lab Invest 1999;59:487-90.
44. Harris EK. Statistical principles underlying analytic goal-setting in clinical chemistry. Am J Clin Pathol 1979;72:374-82.
45. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. Clin Chem Lab Med 2015;53(6):829-32.
46. International Organization for Standardization (ISO). Medical laboratories - particular requirements for quality and competence ISO 15189:2012. Geneva, Switzerland; 2012.
47. Guzel O, Guner EI. ISO 15189 accreditation: Requirements for quality and competence of medical laboratories, experience of a laboratory I. Clin Biochem 2009;42(4-5):274-8.
48. Hovind H, Magnusson B, Krysell M, Lund U, Mäkinen I. Internal quality control - Handbook for chemical laboratories. Nordic Innovation. 2011.
49. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232-5.

Aims and Outline

As shown in the introduction, quality assurance in the clinical laboratory is essential to guarantee appropriate interpretation of clinical IVD test results. Laboratories strive to ensure that correct results are obtained for the samples they analyze to prevent negative impact on the patient's health. To accomplish this, QC systems, such as IQC and EQA, are implemented by the clinical laboratory community. These are not only used to identify (the cause of) erroneous results, but also to provide a means to correct for the observed errors. Both IQC and EQA have a prominent position in quality assurance, however, these tools have limitations. Among others, the artificial nature of QC materials (e.g. pooled, stripped, dialyzed, delipidated, fortified, lyophilized…) can hamper correct interpretation of analytical errors, because the materials don't necessarily reflect the reality of performance of IVD tests on patient samples. As mentioned in the introduction (on page 16), these problems are referred to as (non-)commutability issues. Therefore, IQC and EQA can sometimes fail to identify or falsely suggest significant analytical variation/error at the patient level. In addition, because of the aforementioned potential non-commutability, comparison of QC data can only be reliably done at the peer group level, which makes it more difficult to compare quality across assays. Therefore, this thesis introduces a number of alternative quality management tools which can be used together with the established QC system to aid in providing global evidence about field performance of IVD tests.

The tools described here are intended to monitor the analytical quality of clinical IVD tests by using patient results instead of QC materials, simply because "only good samples can make good assays" (1). Therefore, they essentially avoid problems related to non-commutability. The studies and surveys performed with these tools are further designed to establish a bottom-up approach to quality assessment which facilitates collaboration between laboratories and IVD manufacturers. The aim is to offer involved parties evidence on quality of performance on patient samples, better data access, and a communication platform. To do so, this thesis project is divided in two big parts. The first part involves stability monitoring (chapter 1, 2 and 3), whilst the second focuses on the comparability status of the different IVD tests on the market (chapter 4).

Chapter 1 focuses on the traditional stability assessment by means of IQC. It discusses what type of stability information can be extracted when QC materials are evaluated on either a daily or a monthly basis. To this end, a number of datasets are processed to distinguish between shifts caused by container- and/or lot-effects and calibration aberrations. **Chapter 2** and **chapter 3** illustrate two quality management tools, called the Percentiler and the Flagger, which can be used for mid- to long-term quality monitoring. With these applications one can perform continuous stability performance for 22 common analytes by calculating their daily (out)patient medians (for the Percentiler) and hypo-, and hyper-flagging rate (for the Flagger), which both have to be sent to the project's database. These two web applications allow data visualization based on graphical presentation of the moving medians in time and summary statistics. Thereby, they provide evidence about stability of performance and help to elucidate the reasons for assay variation. For example, allocating assay instabilities to either manufacturer-related problems such as lot inconsistencies or laboratory-related problems such as recalibration issues or poor pre-analytics. In addition these chapters describe how these tools can help with the establishment of realistic quality specifications (based on state-of-the-art performance of the assays currently on the market), and how they provide a means for comparison across manufacturers. Furthermore, by comparing an assay's stability in the Percentiler with the stability in the Flagger, these tools strive to translate the effect of instabilities on "surrogate" medical decisions (with the latter being the hypo- and hyper-flagging rates). In **chapter 4**, the subject is changed to trueness and comparability assessment of several IVD instruments by means of the Master Comparison surveys. These studies are designed to provide the participating laboratories with a panel of 20 commutable single donation serum samples to examine for eight different analytes. Thereby, these surveys are created to provide correct information on the comparability and standardization status of assays and laboratories, and to set benchmarks for the intrinsic quality of commercial assays and for laboratory performance. In addition, these studies intend to add information about the reasons for assay bias (laboratory or manufacturer performance). Next, **chapter 5** describes the communication platform, called the Empower project. It is the overarching concept which encompasses, amongst others, the Percentiler, the Flagger and the Master Comparison surveys. Overall, the Empower project intends to promote the quality, stability, and comparability of global IVD testing by "bottom-up" cooperation of laboratories and manufacturers. **Chapter 6**, finally, elaborates on some statistical tools described in the guidelines from the Clinical & Laboratory Standards Institute (CLSI). This part of the thesis stresses the need for the correct use of statistics for the verification of reference intervals in the clinical laboratory and for commutability assessment of reference materials used by manufacturers to calibrate their assay.

Note, this thesis is a continuation on the work of Dr. Hedwig Stepman (Analytical quality and its effect on medical decisions of laboratory tests for the management of chronic kidney disease, 2). Whilst her work focused on pilot experiments of different aspects of the Empower project, this thesis builds upon those bricks to guide the Empower project from a pilot to an operational scale. This was accomplished by recruiting laboratories for real-time participation to the project. It is also a joint cooperation with the thesis of Linde De Grande. While her project focuses more on the application of the Percentiler and Flagger in the standardization and/or harmonization of thyroid hormone measurements, this thesis emphasizes on clinical chemistry analytes.

References

1. Van Houcke SK, Thienpont LM. "Good samples make good assays" – the problem of sourcing clinical samples for a standardization project. Clin Chem Lab Med 2013;51(5):967-72.
2. Stepman HC. Analytical quality and its effect on medical decisions of laboratory tests for the management of chronic kidney disease – calcium, albumin, phosphorus, 25-hydroxyvitamin D and parathyroid hormone. Doctoral Dissertation 2013.

1 Chapter I

Internal Quality Control monitoring

Highlights

Internal Quality Control (IQC) plays a key role in ensuring reliable results for patient samples analyzed in the clinical laboratory. It is based on the analysis of materials which act as surrogate for patient samples, and is intended for monitoring of the day-to-day precision and accuracy of a laboratory. Different control materials can be used and measured on a daily basis, however, data interpretation can be done at different frequencies. Here, we discuss interpretation of IQC data either on a daily or a monthly basis. We describe what type of information can be extracted from these stability studies.

Highlights of the research:

- Daily IQC results are ideal to focus on short term (in)stability effects, however, they can also be used to address mid- to long-term effects.
- Monthly IQC results can only address long-term effects.
- IQC data can make it possible to distinguish between shifts caused by container- and/or lot-effects and calibration aberrations.

1. Introduction

IQC is the most important and most appropriate tool in a clinical laboratory to monitor the day-to-day precision and accuracy of the applied assays (1). The major objective of (I)QC is to identify measurements where significant errors occurred as well as the reason of those errors, and to stop the process before it exceeds a critical limit of deterioration (2). It thus intends to contribute to the quality of clinical data by guaranteeing reliable laboratory results so that an appropriate clinical interpretation can be made.

To perform IQC in the best possible manner, the laboratory should use IQC materials which fulfill a number of requirements. The applied quality materials should have clinically relevant concentrations of the analyte (preferably at decision limits), and they should be treated in the same manner as the patient samples so that every key step in the process is covered. In addition, they should resemble patient samples as closely as possible, so that any error observed for the QC materials also applies to the patient samples. However, in practice, this can be difficult to achieve due to non-commutability issues and potential instability of the QC material (3,4).

The commercially available IQC materials can be bought from two different sources. In some cases manufacturers provide control materials for their own assays. These types of QC materials are often manufactured from the same material as the calibrator (3). However, the approved CLSI guidelines for statistical QC state that QC materials should be different from the calibrator materials to ensure that the QC procedure provides an independent assessment of the measurement procedure's performance in its entirety, including the procedure for calibration of the measurement (5). It is, therefore, recommended to apply control materials from a third party (such as Bio-Rad or Randox), because they deliver independent assessment of the assays.

In this chapter we interpret a dataset obtained from IQC materials from both sources (manufacturer and third party control). Some of the data were already gathered in the past, but the first analysis was done currently. Unprocessed IQC data which were measured on a daily basis were provided by Roche Belgium (6), in consent with the laboratories which owned the results. Bio-Rad Laboratories (7) provided a second dataset, but this only contained the monthly mean of IQC samples which were also measured on a daily basis. The focus in this chapter lies on the stability observations one can make on the basis of IQC data which are evaluated on a daily (Roche) and on a monthly basis (Bio-Rad). In addition, for the dataset provided by Roche, we try to distinguish between effects caused by mathematical calibration and effects caused by using a new reagent container.

2. Stability assessment based on daily IQC results

A large dataset was received from Roche laboratories (6) comprising data from IQC material which were measured on a daily basis in 43 different laboratories and covering four analytes: calcium, sodium, inorganic phosphor, and albumin. The measurements were done on a Roche Cobas 6000 system and covered a period of 21 months or 639 days starting at 01/01/2011 and ending at 30/09/2012. For some periods there were fewer data available. Therefore, only data from 23/05/11 until 22/05/12 were processed. In addition, data were split up according to the concentration level of the IQC sample (level "391" or "392"). Data analysis was performed with the different levels apart. In case an IQC sample was measured more than once on a single day, the daily average was calculated.

In order to interpret the stability and quality of the IQC measurements, data were visualized with Excel using a moving median grouped per 2, 5, 10, or 15 days. Data analysis started with the 20 labs which had the most values for both concentration levels, but none of them provided data on all of the 639 days of the investigated timeline. Laboratories which provided less than 1000 values for both levels together were excluded from further analysis. In addition, it was necessary to do a mathematical normalization of the two QC material lots which were used for calcium and inorganic phosphor. If not, this resulted in a moving median which was obviously going up or down due to the different concentration of the QC material. The result was a number of graphs presenting each the stability of the IQC measurements of one laboratory in time. Furthermore, by calculating the daily median for all laboratories using the same assay, we could also plot and interpret the stability of the entire Roche Cobas peer group and compare each laboratory with its peer. Around the moving median of the peer group we plotted QC limits which take into account both the biological variation of the studied analytes, and the current capabilities of the applied Roche Cobas assays (calcium: ±0.05 mmol/L; sodium: ±1 mmol/L; inorganic phosphor: ±0.04 mmol/L; albumin: ±1 g/L). We interpreted the quality of the IQC measurements versus the chosen analyte-specific QC limits, but it should be noted that a consensus on these does not exist (8).

Depending on the applied value for the moving median, the Excel graphs can either be used to detect short term quality problems (*n* = 2, 5) or they can be applied for mid-to-long term monitoring and interpretation of the analytical quality (*n* = 10, 15). In Figure 1 an example can be seen for lab 0611-06 for calcium level "391" demonstrating the difference when applying a moving median grouped per 2, 5, 10, or 15 days. The IQC data from this laboratory (indicated in orange) shows both good comparability to its peer group (indicated in red) and sufficient stability within the applied QC limits (indicated in black; around the peer group median).

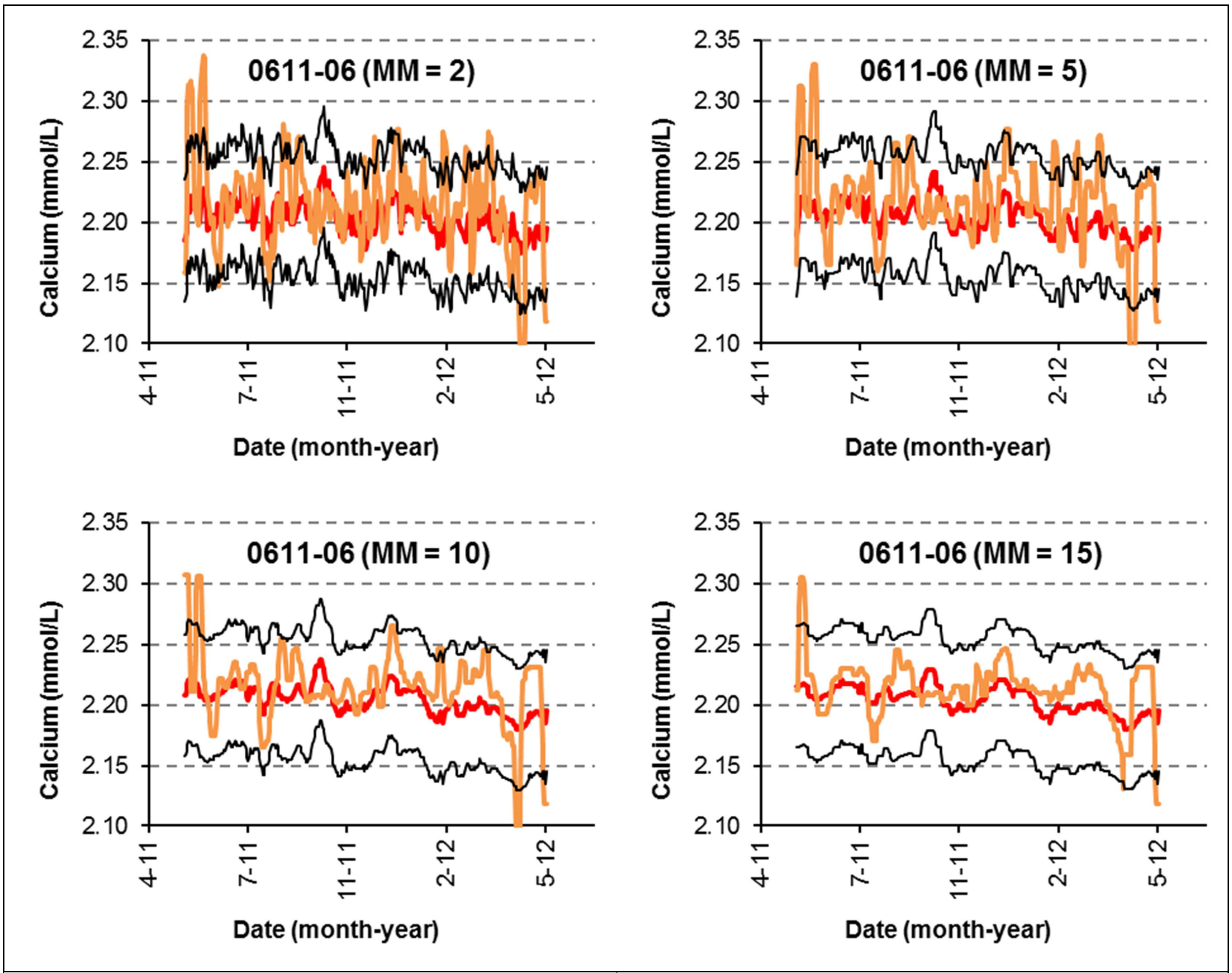


Figure 1. Calcium IQC data for one laboratory 0611-06. Orange lines represent the moving median (MM; n = 2, 5, 10 or 15 days) for the individual laboratory. The red lines correspond with the moving median of the Roche Cobas peer group, the black lines with the limits (±0.05 mmol/L for calcium).

Described on the next four pages are four examples for each analyte, but only data on the “391” level are discussed (Figures 2-5). A complete overview of all data can be found in the annex (Supplemental Figure 1-4). Note, only moving medians for *n* = 10 are shown. This is a good approach to identify general effects, but one can miss short term stability/quality problems. For the latter, a moving median of 2 or 5 should be used or the data should be plotted with a scatterplot instead of using a moving median.

Albumin (see Figure 2)

- 0807-03: a calibration aberration which occurred in fall 2011 led to an upwards shift borderline to significant outside of the defined QC limits.
- 0969-21: long-term stability of the laboratory is interrupted by a downwards drift (starting from 2012), leading to a (borderline) negative bias.
- 0952-15: the laboratory starts with a number of low values for the IQC material, but improves and has generally good stability at the end of the timeframe.
- 0616-06: laboratory with a good stability. Only two shifts possibly occurred over the complete timeframe, but these didn't lead to a significant bias compared to the peer group.

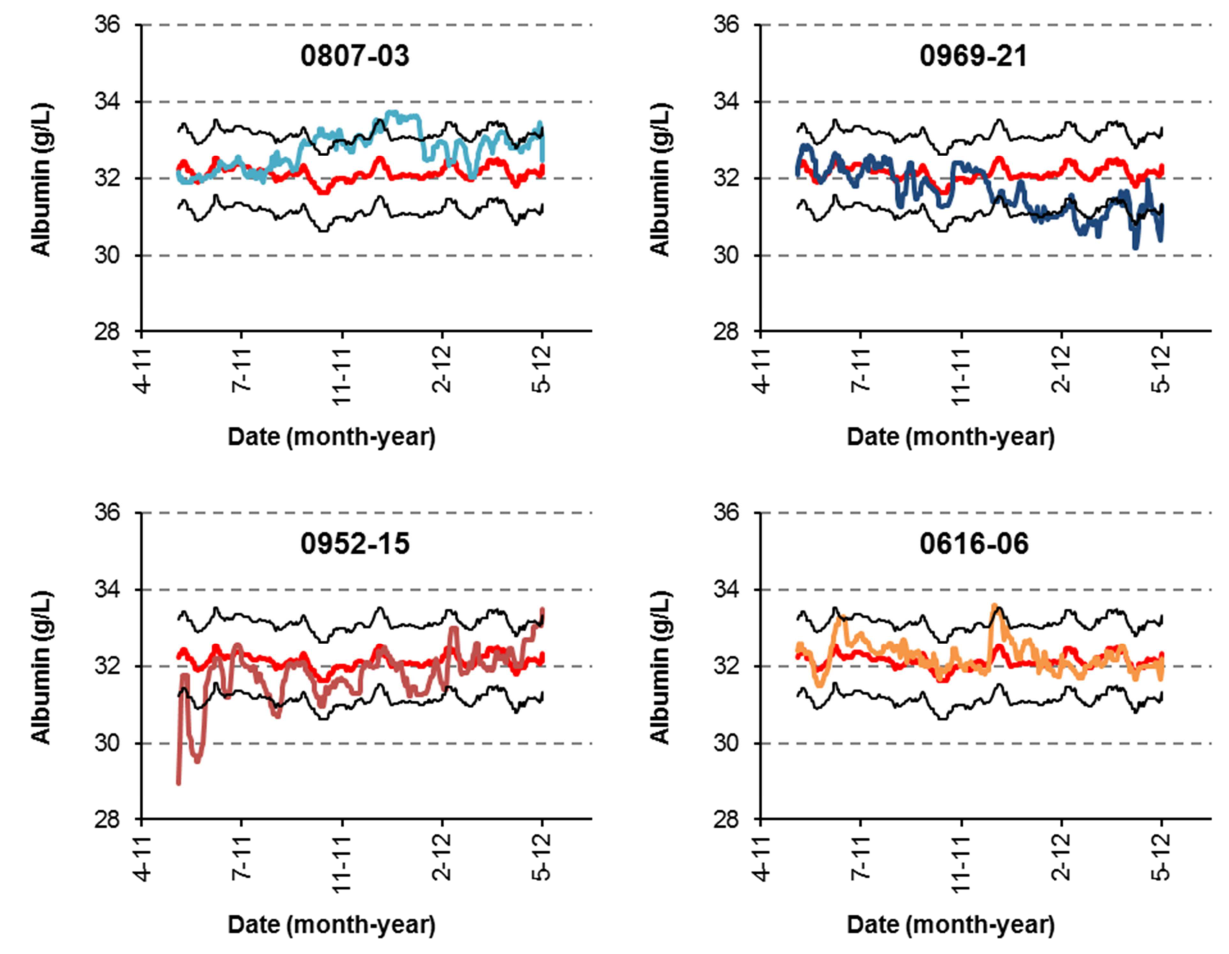


Figure 2. Albumin IQC data from Roche for four laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied QC limits (± 1 g/L) are shown in black lines.

Calcium (see Figure 3)

- 1098-13: high variation of the measurements of the IQC material. The calcium assay of the laboratory shows a sawtooth pattern, indicating a drift caused by an unstable assay, which is followed by a recalibration. At the end of the timeframe a positive bias arises.
- 0969-21: a long term negative bias from mid to the end of 2011. Comparability to the peer target improves in 2012, but the data still suffer from a high variation.
- 0625-01: a very clear shift occurs near the end of 2011 moving the data from a bias which is borderline positive to a borderline negative bias.
- 0765-05: a long term borderline negative bias within the QC limits.

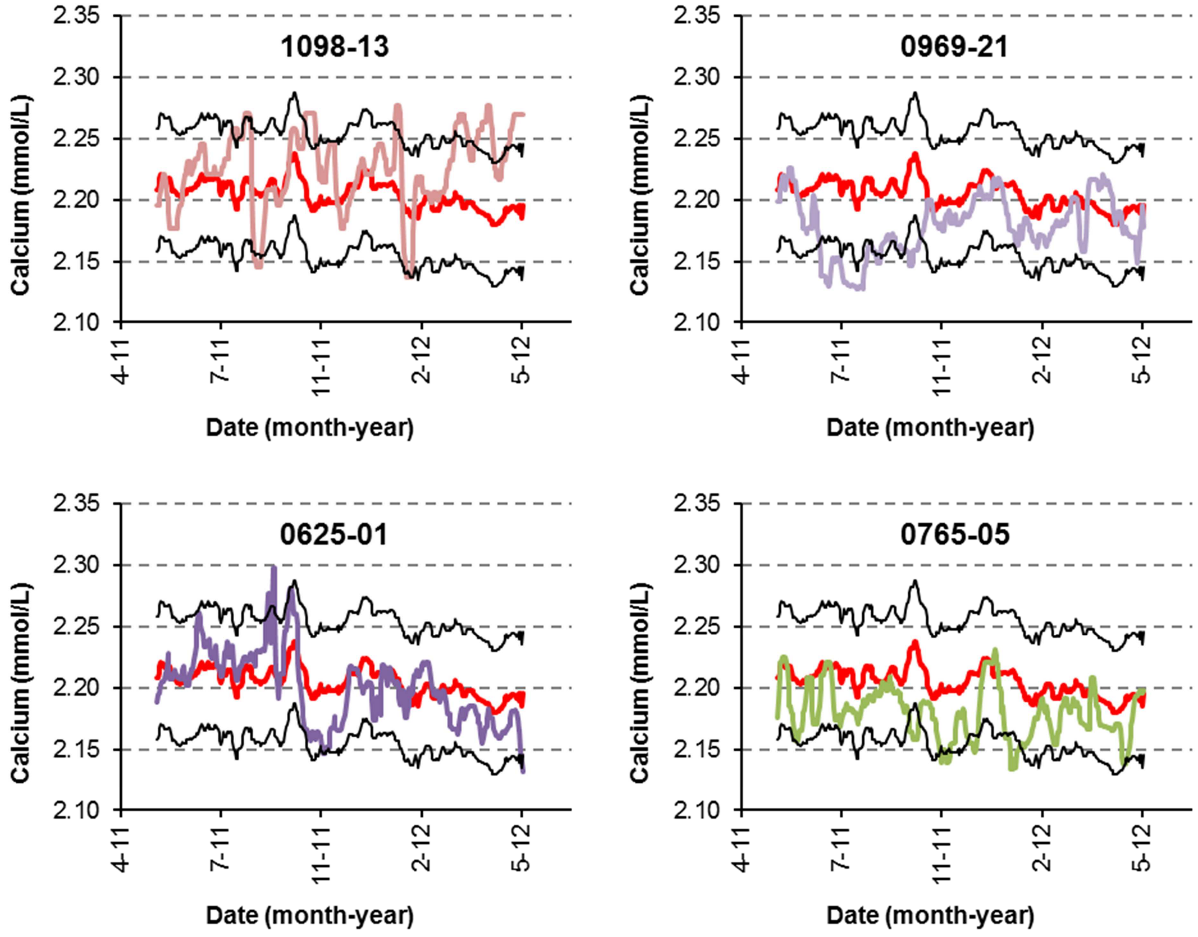


Figure 3. Calcium IQC data from Roche for four laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied QC limits (± 0.05 mmol/L) are shown in black lines.

Inorganic phosphor (see Figure 4)

- 1098-13: insignificant negative bias, within the specification limits. Pearls on a string effect caused by rounding of the result.
- 0954-25: data was mostly stable with the exception of a very clear calibration aberration around March 2012.
- 0811-20: two positive shifts (September – December 2011 and February – April 2012) led to a significant long term positive bias. Each time this effect was cancelled by a downward shift restoring values which are comparable with the peer group median.
- 1006-08: a laboratory with a very good stability and comparability to the peer group.

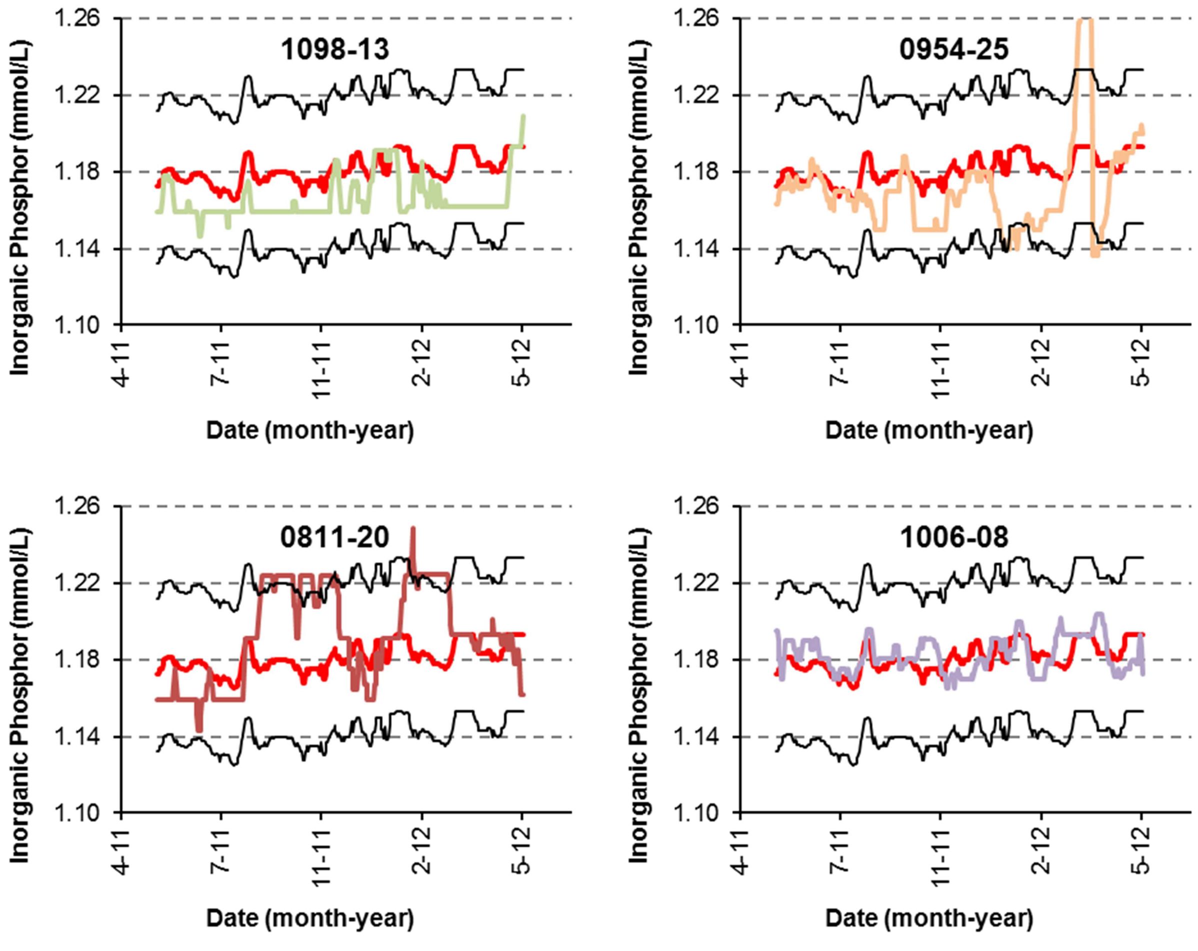


Figure 4. Inorganic phosphor IQC data from Roche for four laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied QC limits (± 0.04 mmol/L) are shown in black lines.

Sodium (see Figure 5)

- 0611-06: occurrence of a high variation due to a significant number of calibration aberrations, and a small positive bias.
- 0969-21: a significant long term positive bias exists. It is remediated from 2012 on.
- 0750-09: a significant long term negative bias occurs which persists over the complete timeframe.
- 0952-15: a laboratory with a very good stability and comparability to the peer group.

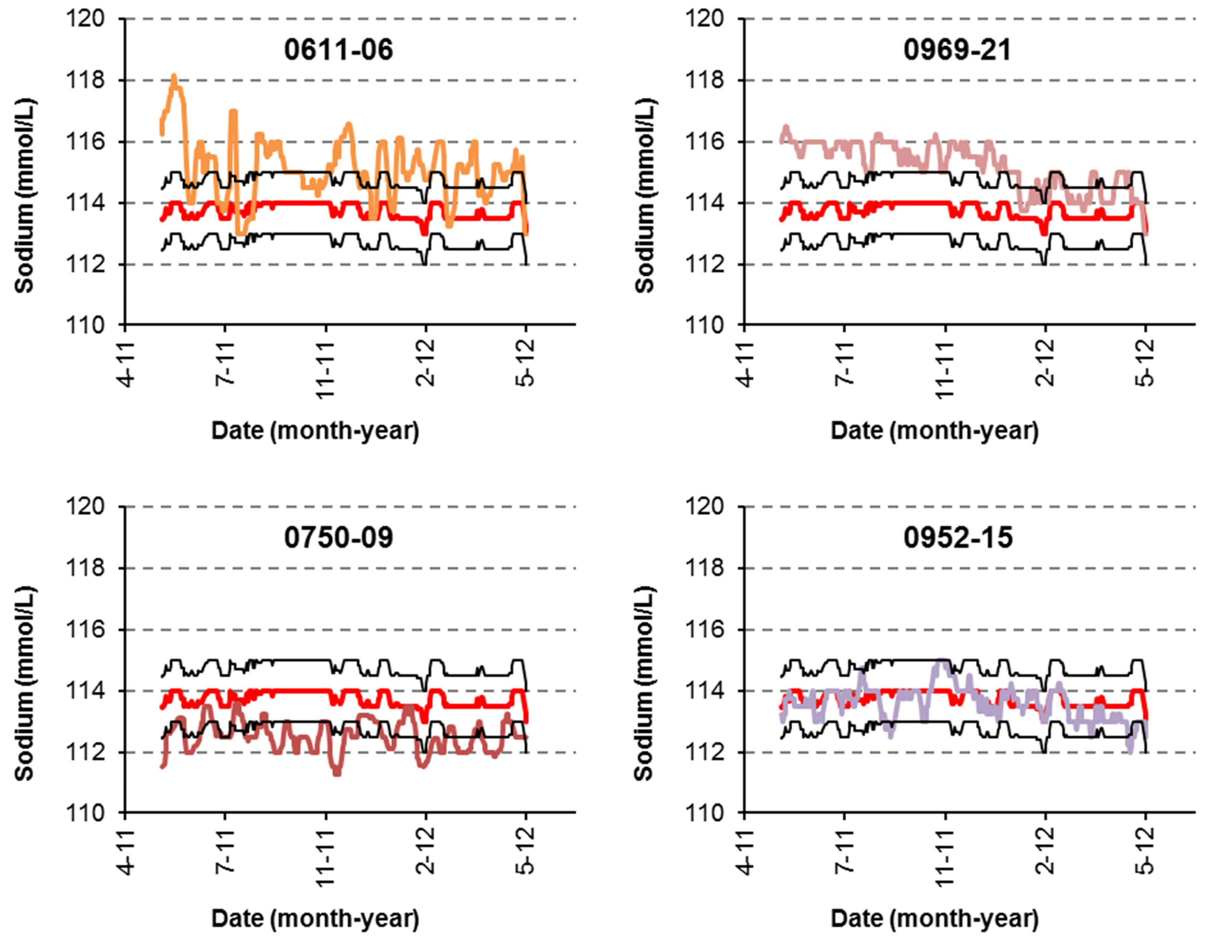


Figure 5. Sodium IQC data from Roche for four laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied QC limits (± 1 mmol/L) are shown in black lines.

3. Container and calibration effects

Roche IQC data for albumin, calcium, and inorganic phosphor, also included the reagent container identification number of each measurement performed. With this information one can distinguish between analytical shifts which are caused by a new reagent container or by mathematical calibration. In theory one can also distinguish shifts which are caused by a new lot, but the lot numbers were not supplied with the data. Each measurement result per day was plotted in a scatterplot. Different reagent container ID’s were visualized with a different symbol. For each analyte, a dataset of two laboratories is shown. Figures 6-7 show that some containers are subject to a higher drift than others, caused by a higher instability. When a shift occurs during the use of a single container, this can be allocated to a calibration effect. An example can be seen in Figure 7 for calcium (middle row, left column, blue diamonds). When a new container is taken into service, a shift can either be caused by the new container itself or by a combination of a new lot and new container. An example can be seen in Figure 7 for inorganic phosphor (bottom row, left column, transition from yellow circles to red squares).

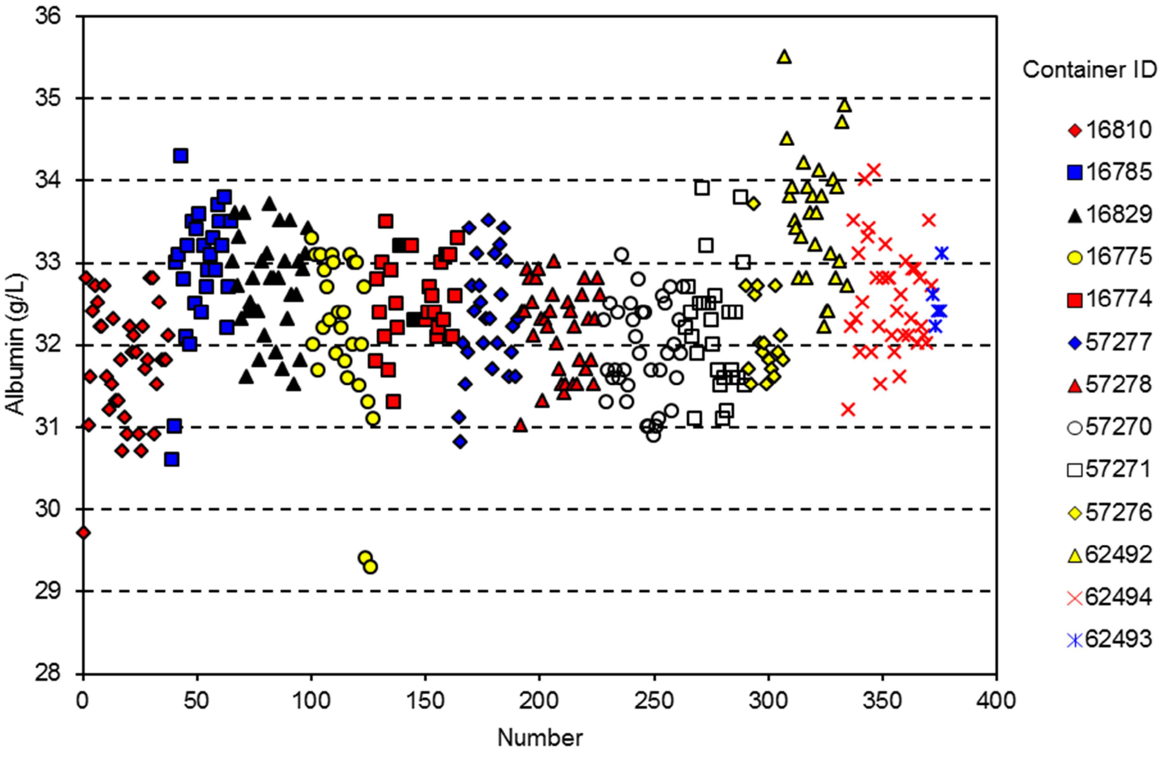


Figure 6. Container and calibration effects for albumin (one laboratory). Different container ID’s are visualized with a different symbol.

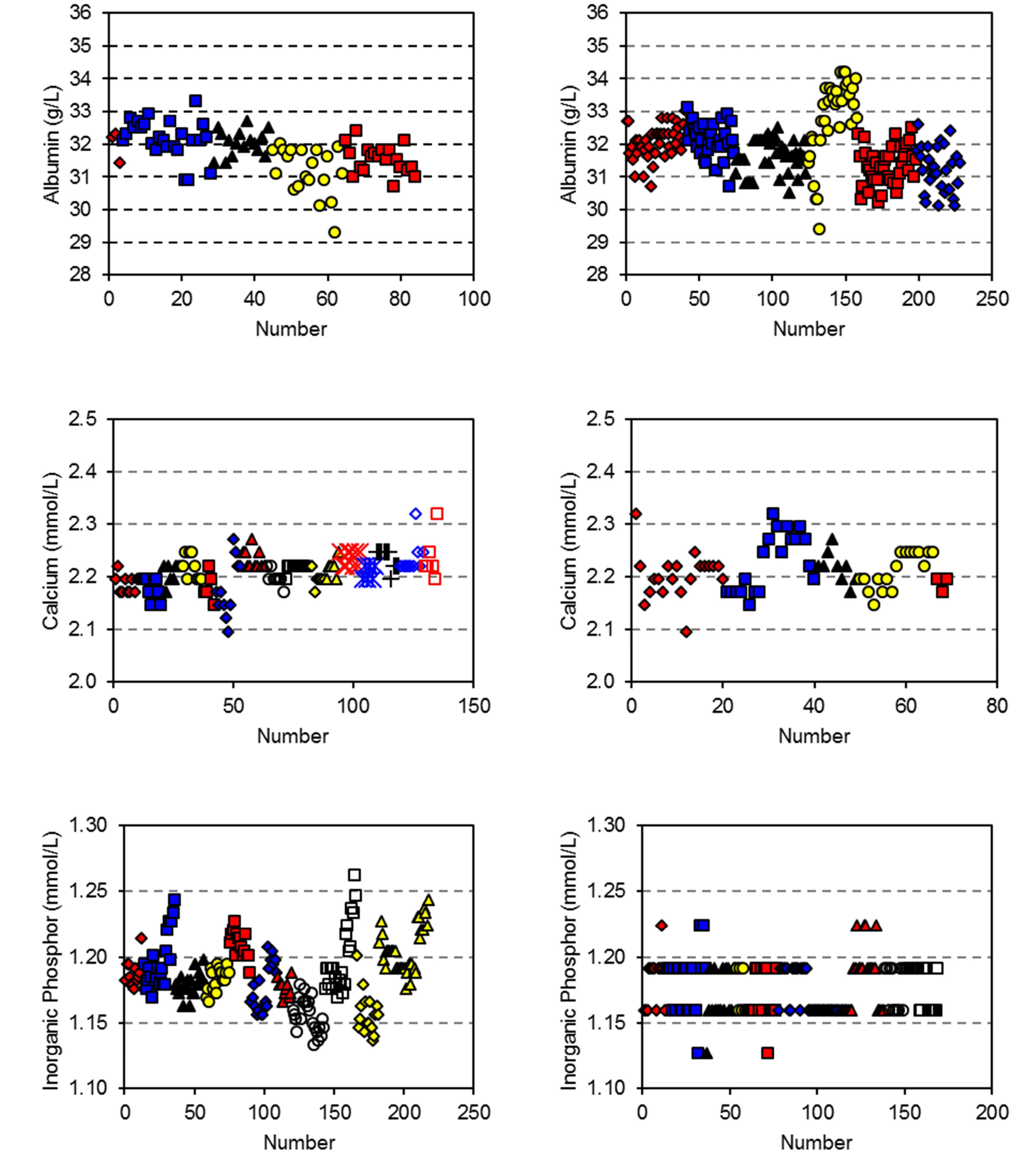


Figure 7. Container and calibration effects for albumin, calcium and inorganic phosphor (two laboratories each). Different container ID’s are visualized with a different symbol.

4. Stability assessment based on monthly IQC results

A second large dataset was received from Bio-Rad Laboratories Inc (7) comprising IQC data from 430 different laboratories and covering the same four analytes: calcium, sodium, inorganic phosphor, and albumin. The IQC measurements were performed daily by the laboratories but only the monthly mean was provided by Bio-Rad. The dataset covered measurements on seven different devices (Abbott Architect, Beckman DxC, Ortho Vitros, Roche Cobas, Roche Integra, Siemens Advia, and Siemens Dimension Vista), and a period of about 20 months going from October 2010 until June 2012. In order to interpret the stability and quality of the IQC measurements, data were visualized with Excel using a scatterplot of the monthly means. Shown and interpreted in this section are data for laboratories using the Roche Integra device. The same QC limits as before, i.e., that take into account biological variation and the current capabilities of the assays, were applied in the graphical representation (calcium: ±0.05 mmol/L; sodium: ±1 mmol/L; inorganic phosphor: ±0.04 mmol/L; albumin: ±1 g/L). Again, these QC limits are plotted around the Roche Integra peer group median values.

With the graphs, we described four example laboratories for each analyte. The other data were added in the annex (Supplemental Figure 5-8). Note, because only monthly means were reported these data can only be used to visualize long term effects. Short term and even mid-to-long term effects can’t be addressed.

Albumin (see Figure 8)

- 155555: laboratory with a very good stability and comparability to the peer group. Only at the end of the timeframe a potential problem occurred.
- 210311: a borderline to significant negative bias compared to the peer target.
- 916217: the laboratory started with a good stability but a shift in mid-2011 caused a significant positive bias.
- 916229: this laboratory shows the same trend as number 916217 and other laboratories (data not shown). This might indicate that a new lot was deployed by these IQC participants. Because the majority of laboratories didn't show this trend, the peer median is not influenced by the potential new lot.

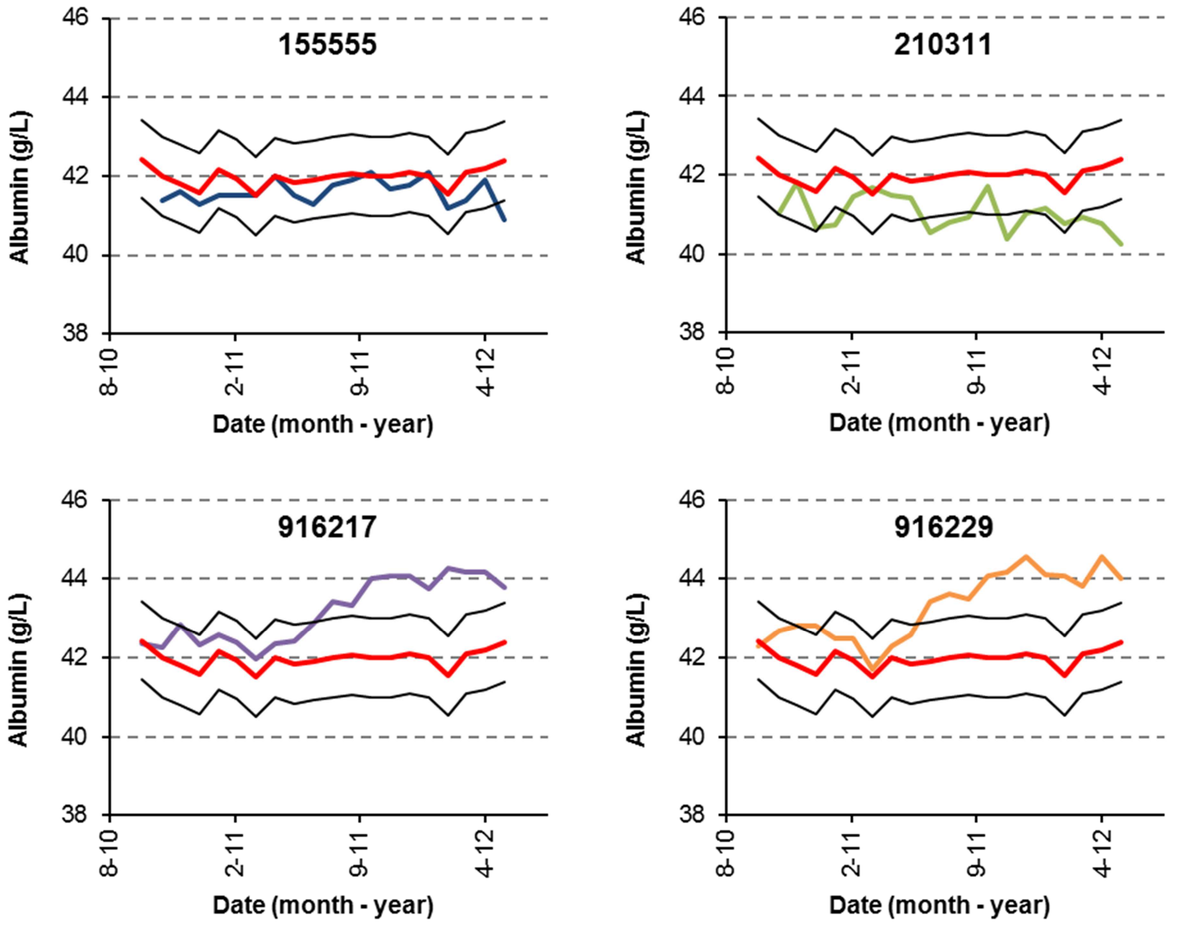


Figure 8. Albumin IQC data from Bio-Rad for four laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 1 g/L) in black lines.

Calcium (see Figure 9)

- 382382: laboratory with a good stability and comparability to the peer group
- 346892: a significant long term positive bias occurs over the complete timeframe.
- 172131: a borderline to significant long term negative bias occurs over the complete timeframe.
- 390916: a borderline to significant long term positive bias occurs over the complete timeframe.

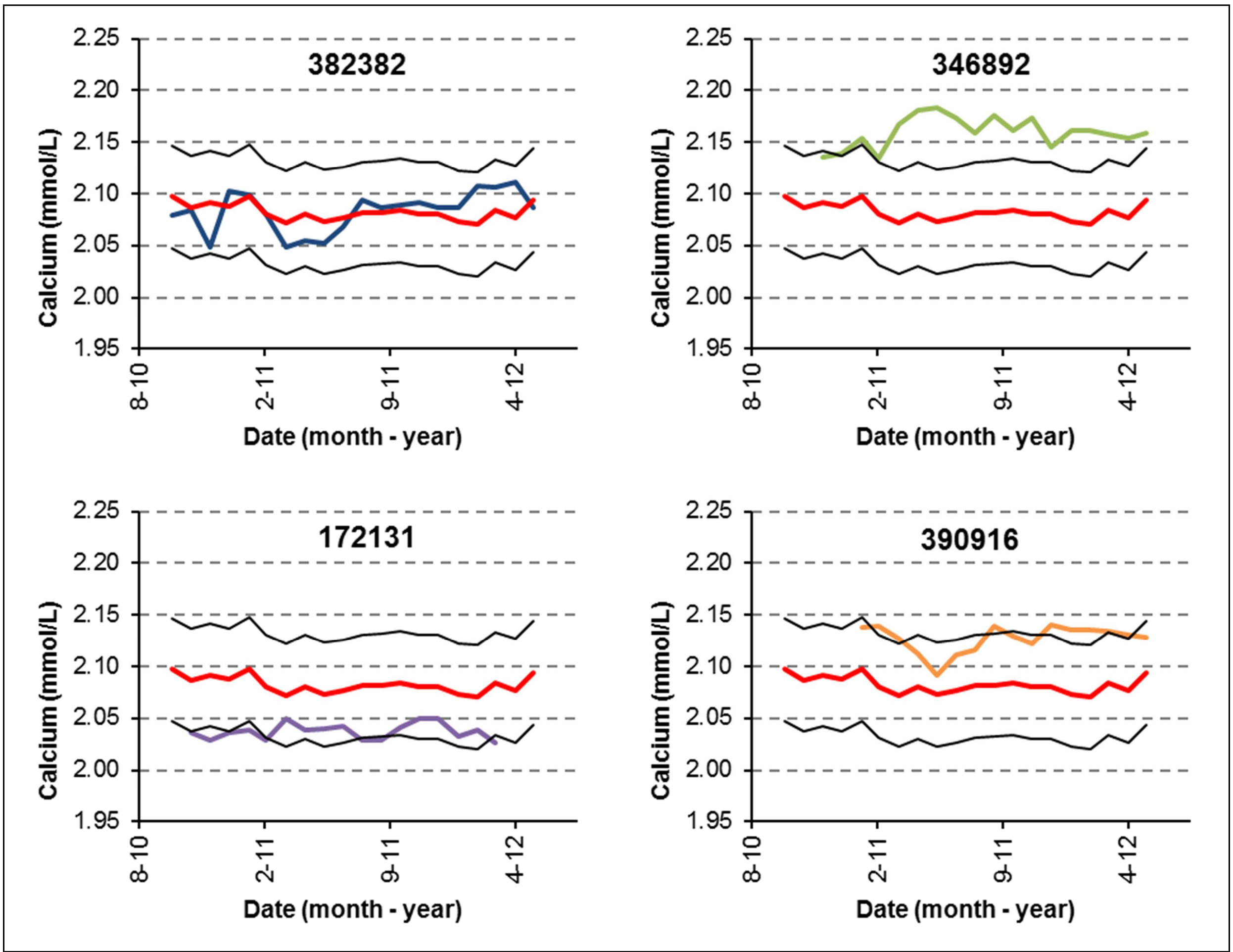


Figure 9. Calcium IQC data from Bio-Rad for four laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 0.05 mmol/L) in black lines.

Inorganic phosphor (see Figure 10)

- 916197: laboratory with a good stability and comparability to the peer group
- 916205: analogous to number 916197, a laboratory with a good stability and comparability to the peer group.
- 346892: laboratory with a good stability but with a long term borderline positive bias compared to the peer group.
- 172131: compared to the other laboratories, a high variation occurs. However, almost all measurement values remain within the quality specifications.

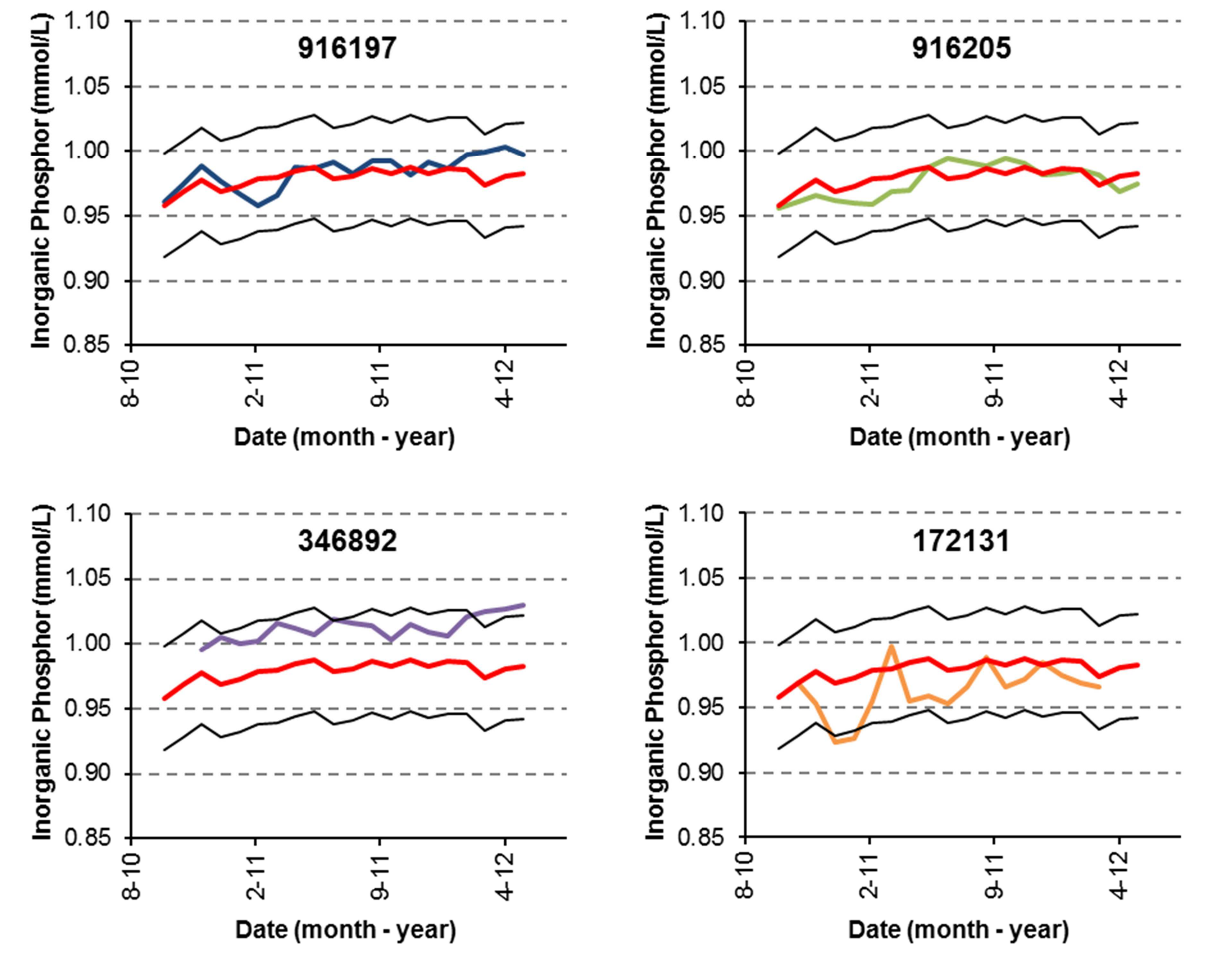


Figure 10. Inorganic phosphor IQC data from Bio-Rad for four laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 0.04 mmol/L) in black lines.

Sodium (see Figure 11)

- 382382: the laboratory started with a good stability but a shift in mid-2011 caused a significant positive bias.
- 916197: this laboratory shows the same trend as number 382382 and other laboratories (data shown in the annex; Supplemental Figure 8). This might indicate that a new lot was deployed by these IQC participants. Because the majority of laboratories didn't show this trend, the peer median is not influenced by the potential new lot.
- 908214: a rather high variation, but most measurements remain within the quality limits.
- 195994: laboratory with a good stability but with a long term borderline negative bias compared to the peer group.

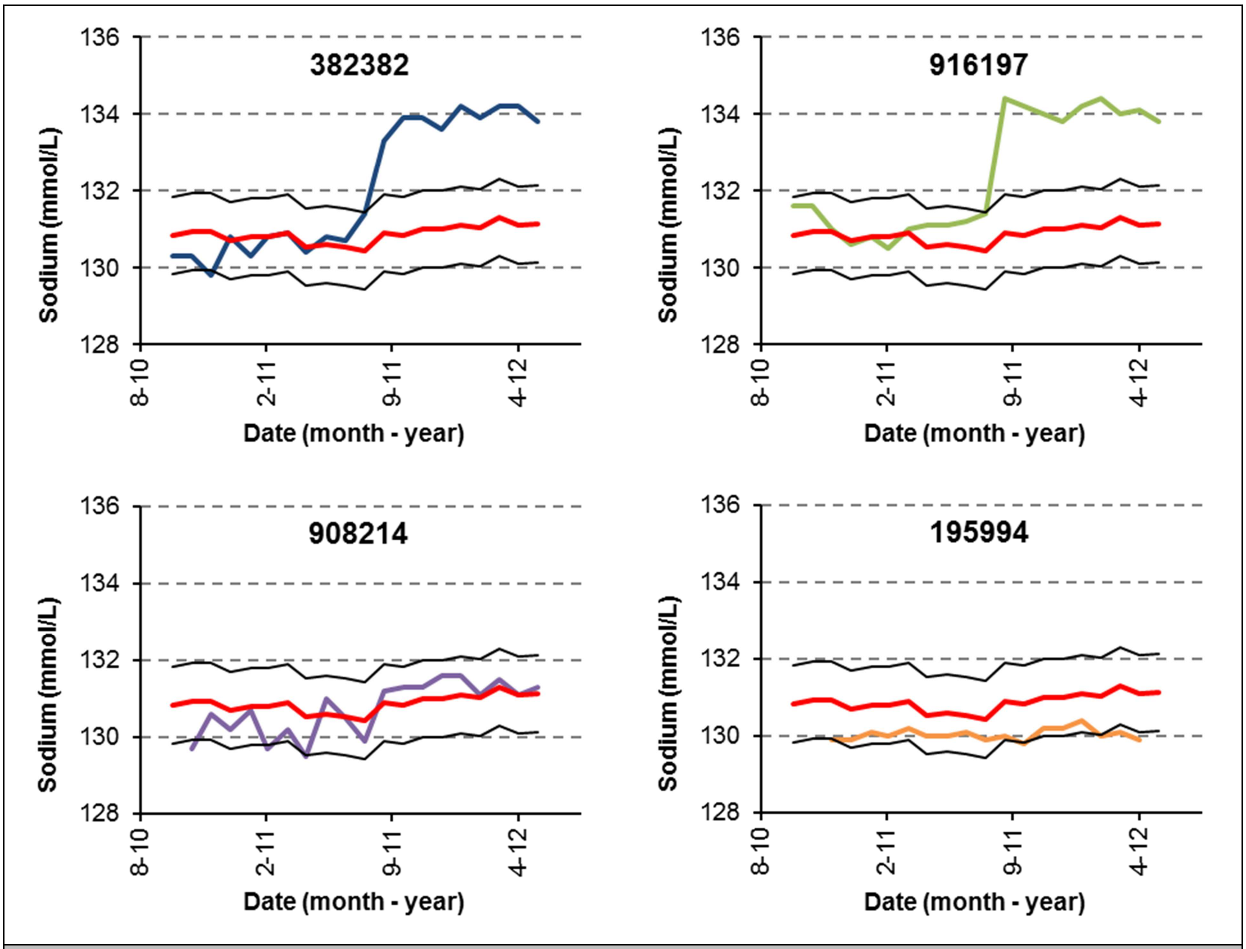


Figure 11. Sodium IQC data from Bio-Rad for four laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 1 mmol/L) in black lines.

5. Conclusion on the use of IQC data

IQC data can provide key insight into a laboratory's performance as the process contributes to a better understanding of analytical error. When performed correctly it allows for early recognition of measurement and technical problems and rapid introduction of counteracting measures (1). Depending on the frequency at which the QC material is evaluated (e.g. daily values grouped per 2, 5, 10, or 15 days or grouped per month), it allows for both short-term and mid- to long-term quality monitoring. However, when the QC material is rarely evaluated (e.g. when using the monthly mean) it loses its ability for early recognition of measurement problems. Therefore, the discussed examples of the Bio-Rad IQC can only be used to get a broad idea about the long-term performance of assays/instruments, whilst IQC data which is interpreted on a daily basis can be used to track effects of unstable assays, (re)calibrations, and the introduction of different lots and/or different reagent containers.

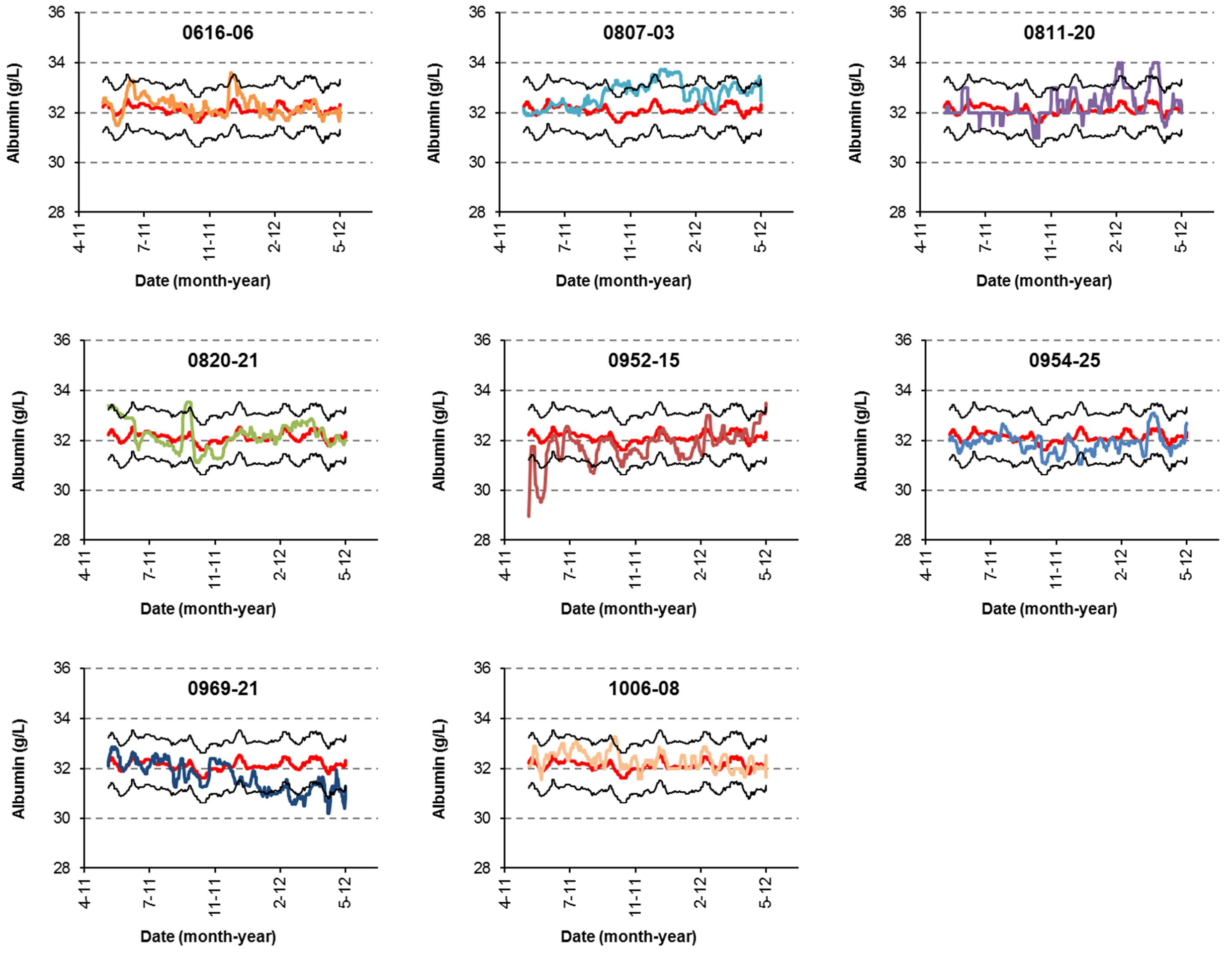
Although data based on QC materials serve as the first line of defense against unreliable patient results, it should not be ignored that, in some cases, trends (shifts or drifts) in measurements of patient samples can be missed by QC materials. For example, commutability problems make them inappropriate to verify consistency of patient sample results when reagent lots are changed. Data from Miller *et al* indicate that there is no reliable relationship between the bias for QC results and the bias for patient results caused by two different reagent lots (4). In addition, these commutability effects make it inappropriate to compare IQC data of laboratories from different peer groups. However, when laboratories use the same instruments and assays, QC materials can still be used to compare the stability and comparability of laboratories with their peer group.

References

1. Lock RJ. My approach to internal quality control in a clinical immunology laboratory. J Clin Pathol. 2006;59:681–4.
2. Badrick T. The Quality Control System. Clin Biochem Rev 2008;29:S67–S70.
3. Kinns H, Pitkin S, Housley D, Freedman DB. Internal quality control: best practice. J Clin Pathol 2013;66:1027-32.
4. Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76-83.
5. Wayne PA. CLSI. Statistical quality control for quantitative measurement procedures: principles and definitions; approved guideline-third edition. CLSI document C24-A3. Clinical and Laboratory Standard Institute; 2006.
6. Roche Belgium. http://www.roche.be/ (accessed January 2015).
7. Bio-Rad laboratories. Discover the importance of third party quality control. USA: Bio-Rad Laboratories 2010 http://www.qcnet.com/thirdpartyQC/PDF/ThirdPartyQC.pdf (accessed January 2015).
8. Cembrowski GS. Thoughts on quality-control systems: a laboratorian's perspective. Clin Chem 1997;43:886-92.

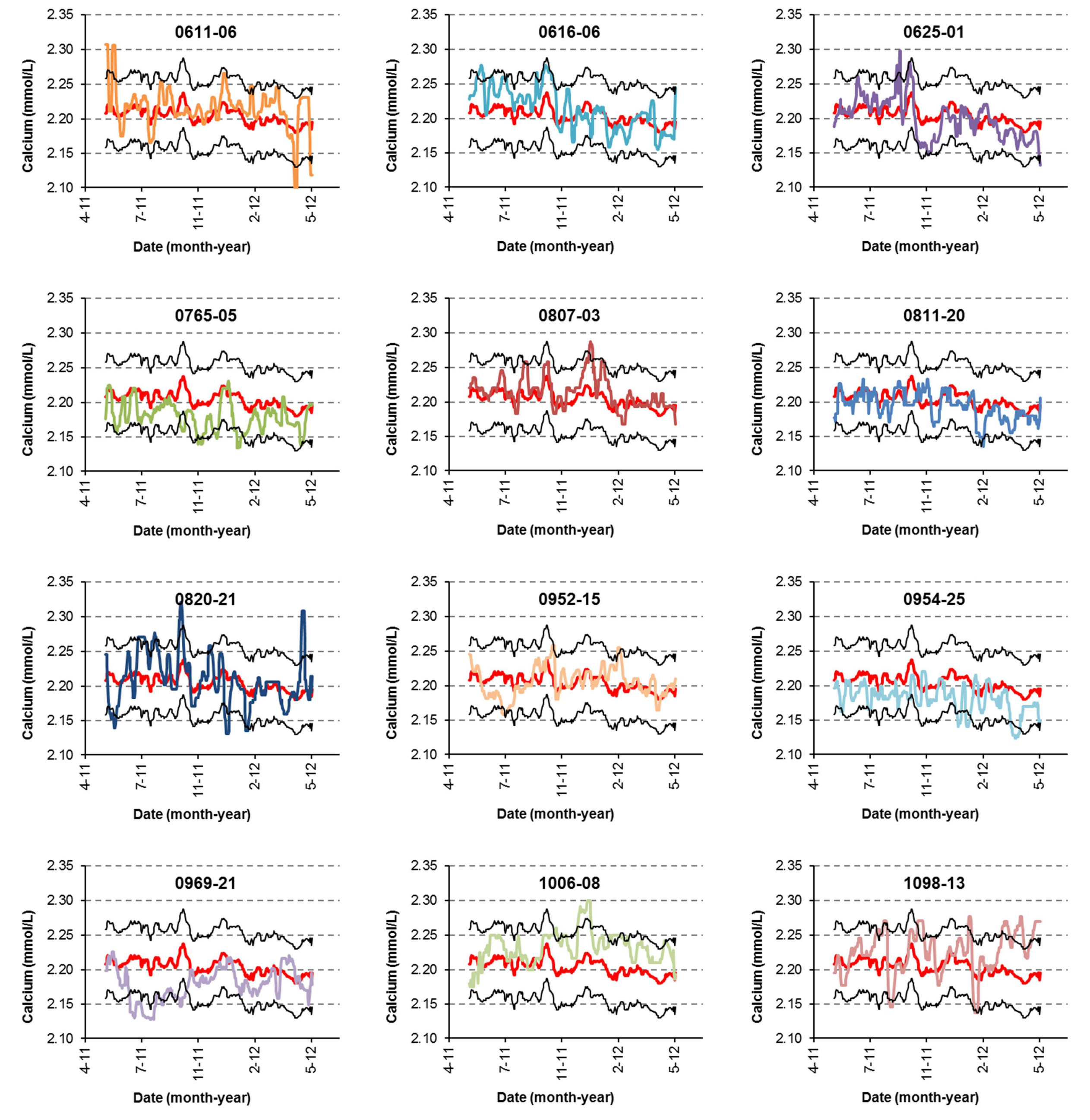
Annex to chapter 1

Roche IQC results for several laboratories on albumin:



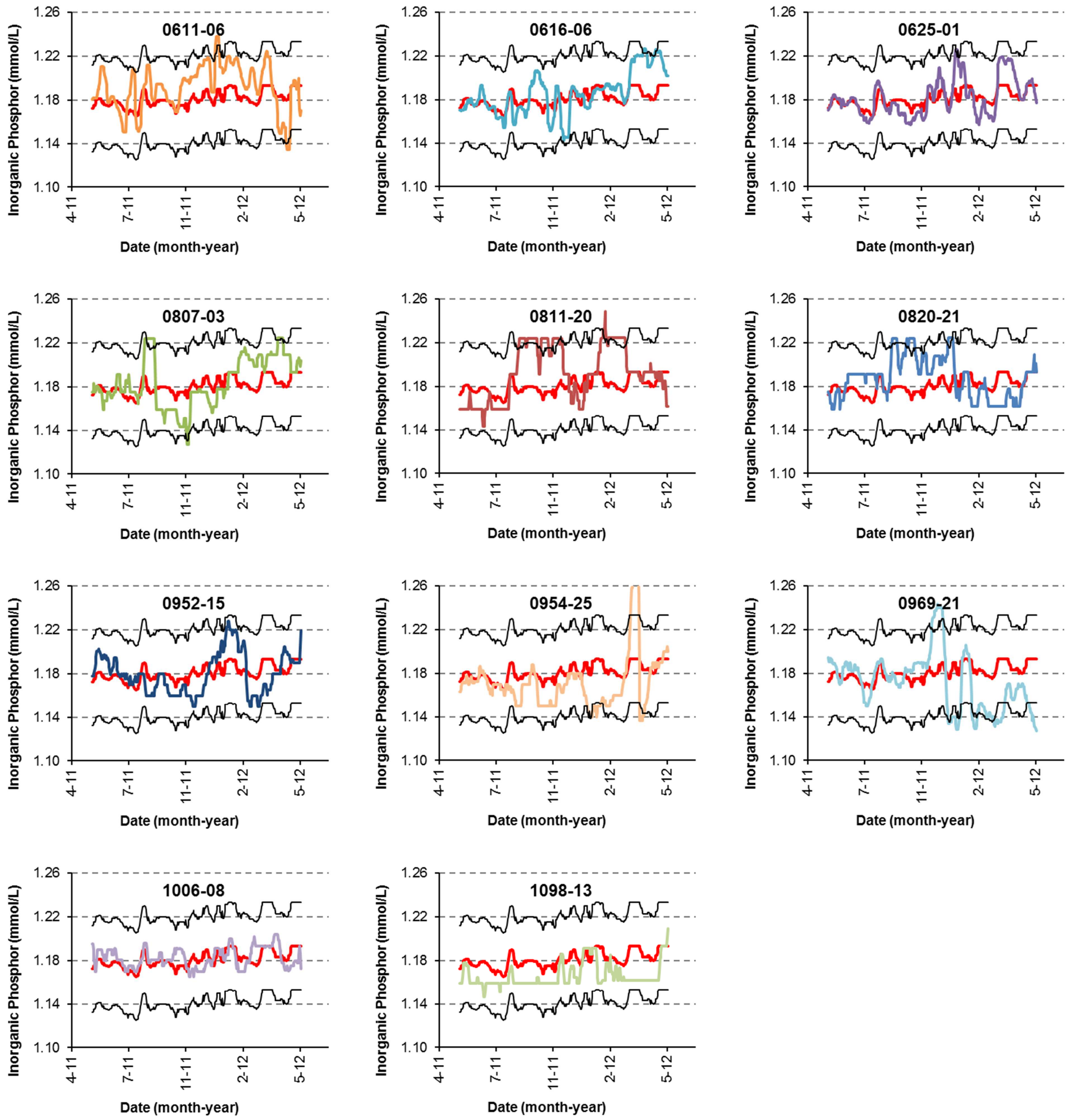
Supplemental Figure 1. Albumin IQC data from Roche for several laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied limits (± 1 g/L) in black lines.

Roche IQC results for several laboratories on calcium:



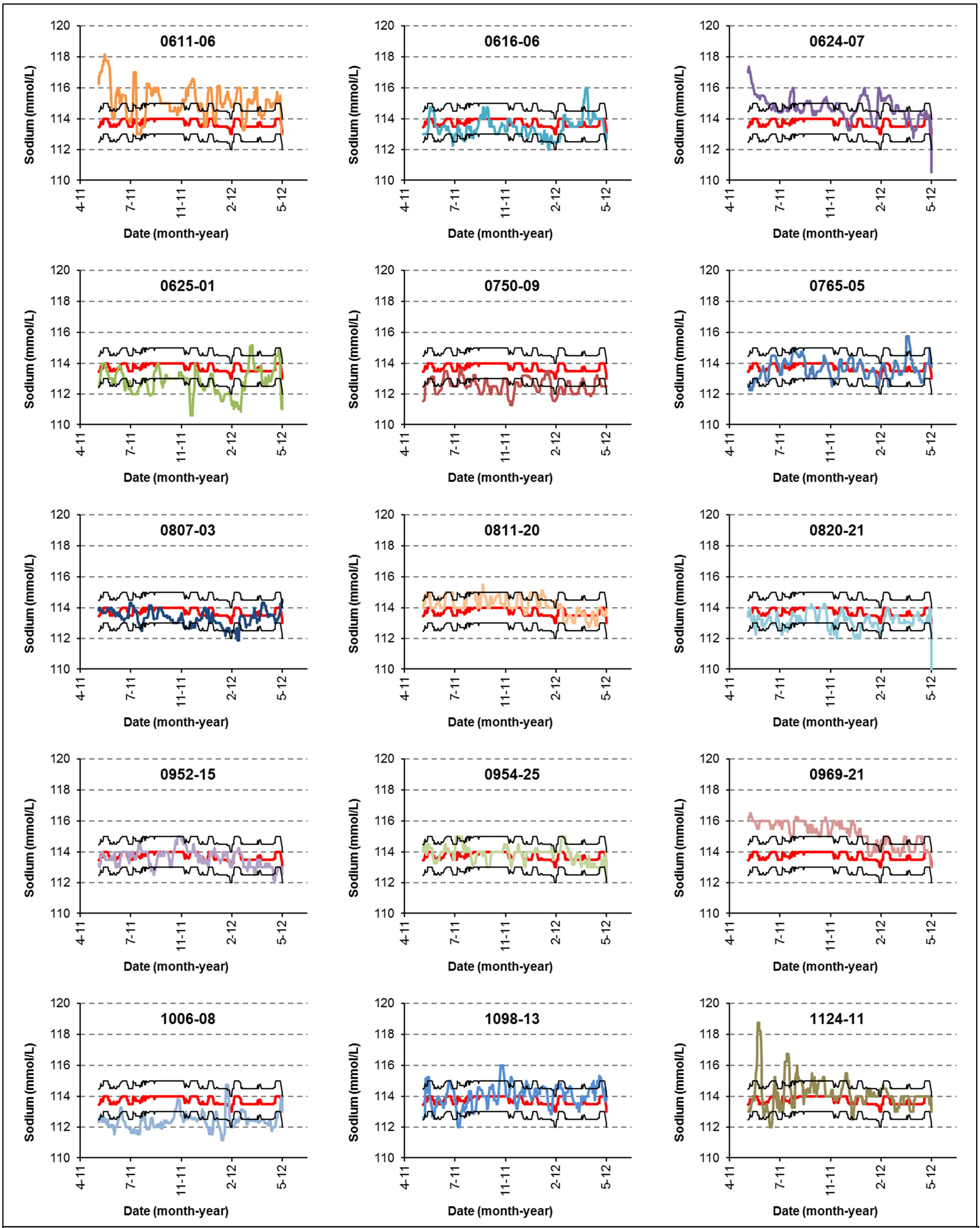
Supplemental Figure 2. Calcium IQC data from Roche for several laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied limits (± 0.05 mmol/L) in black lines.

Roche IQC results for several laboratories on inorganic phosphor:



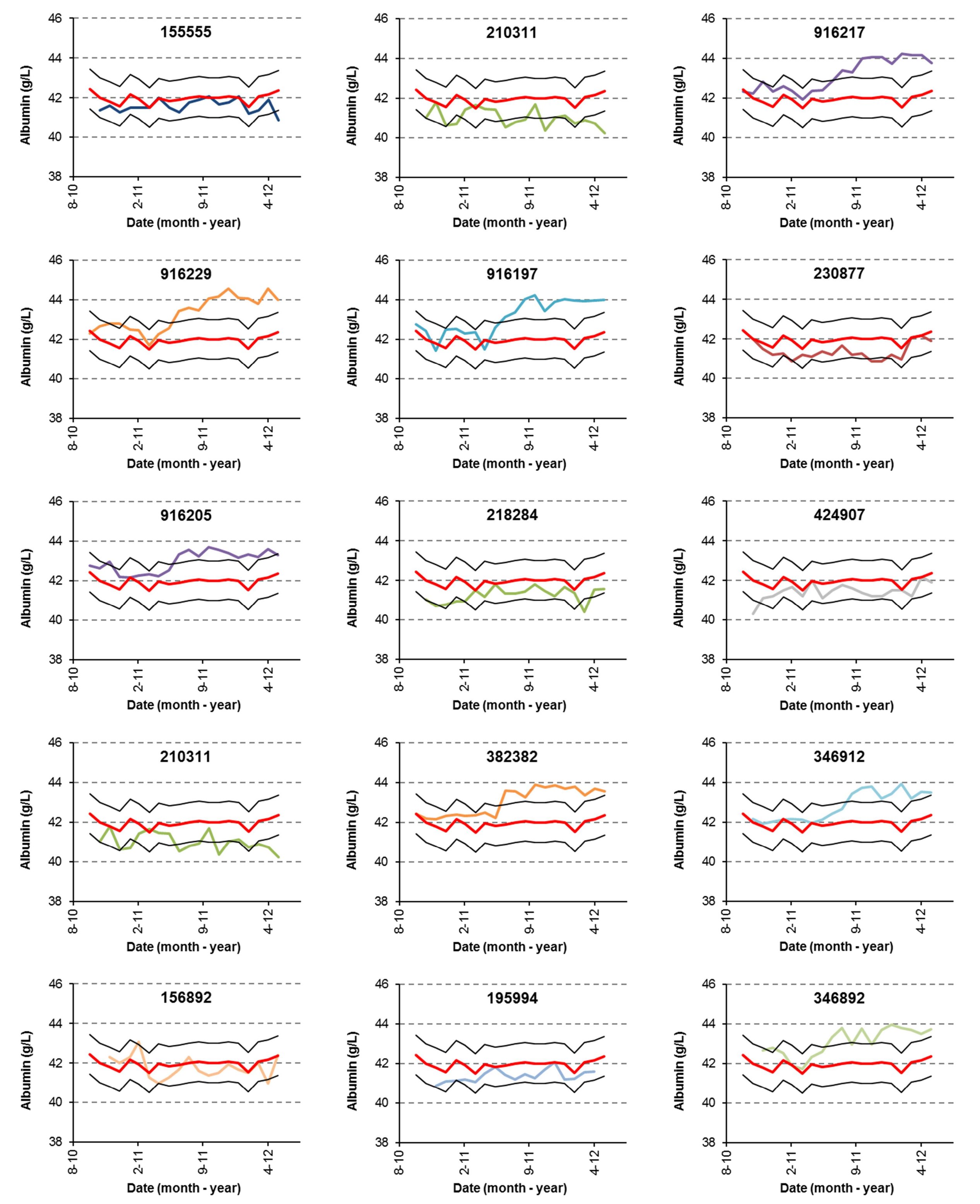
Supplemental Figure 3. Inorganic phoshor IQC data from Roche for several laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied limits (± 0.04 mmol/L) in black lines.

Roche IQC results for several laboratories on sodium:



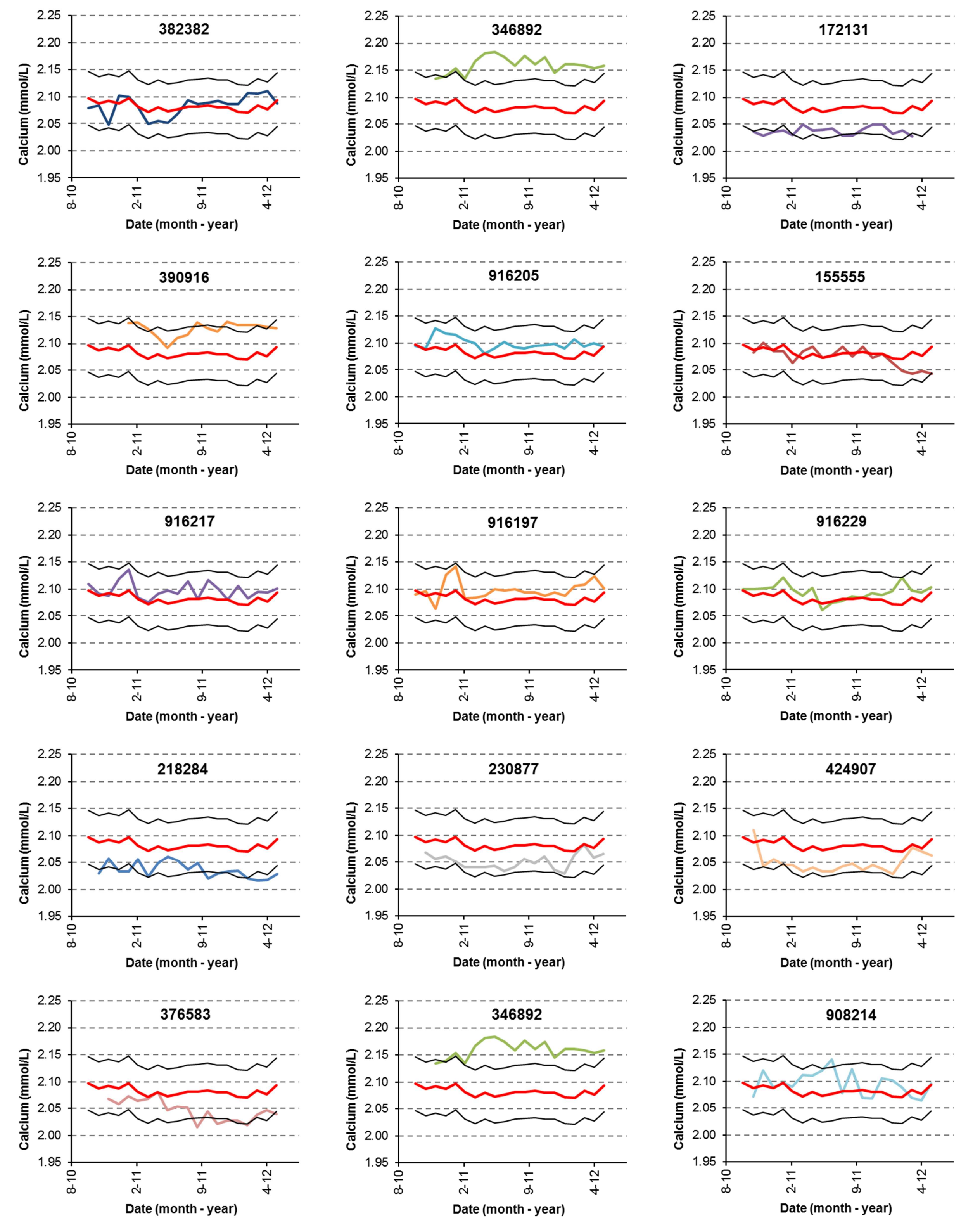
Supplemental Figure 4. Sodium IQC data from Roche for several laboratories. Colored lines (not red) represent the moving median (*n* = 10) of the individual laboratory. The peer group moving median is shown with a red line, the applied limits (± 1 mmol/L) in black lines.

Biorad IQC results for several laboratories on albumin:



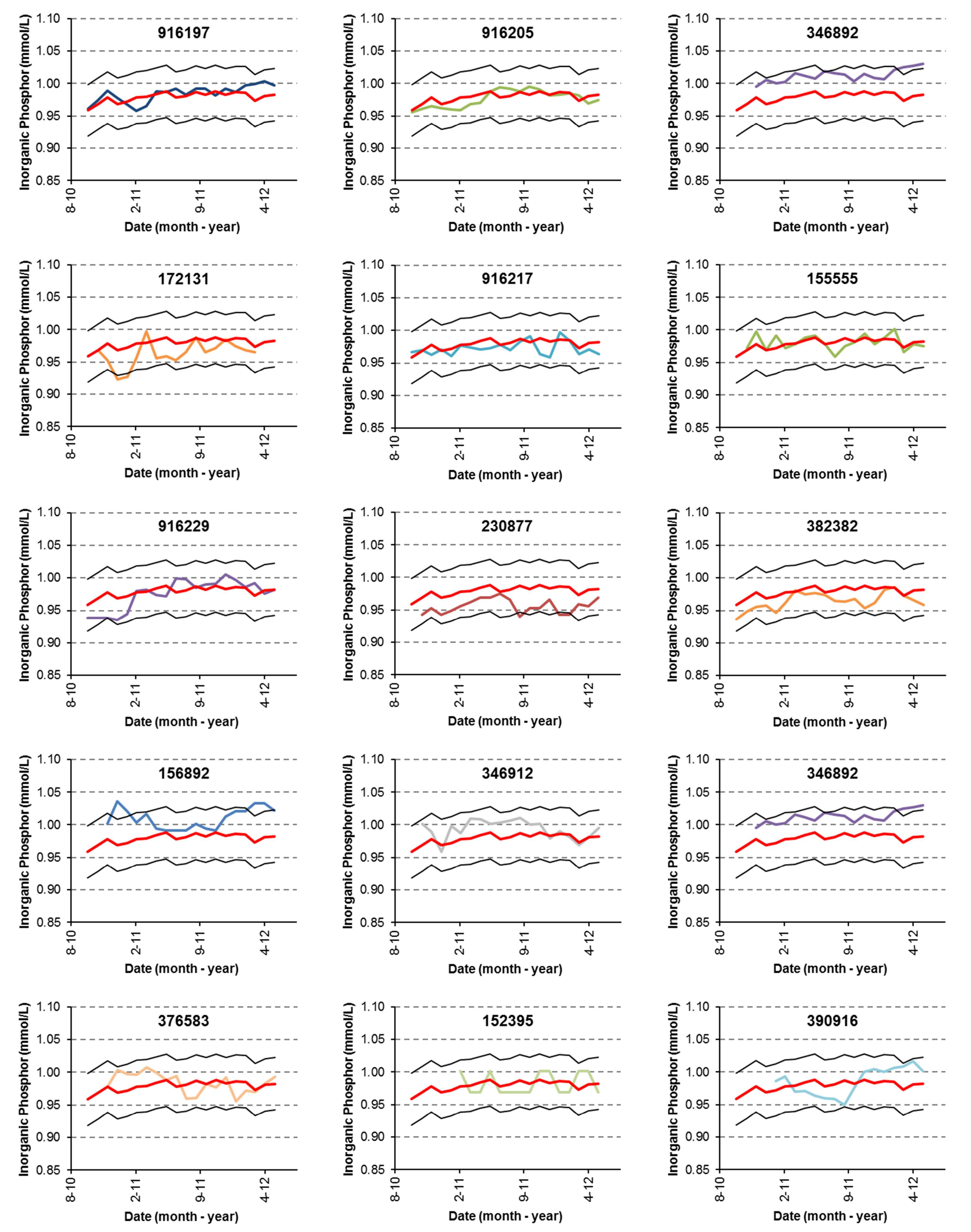
Supplemental Figure 5. Albumin IQC data from Biorad for several laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 1 g/L) in black lines.

Biorad IQC results for several laboratories on calcium:



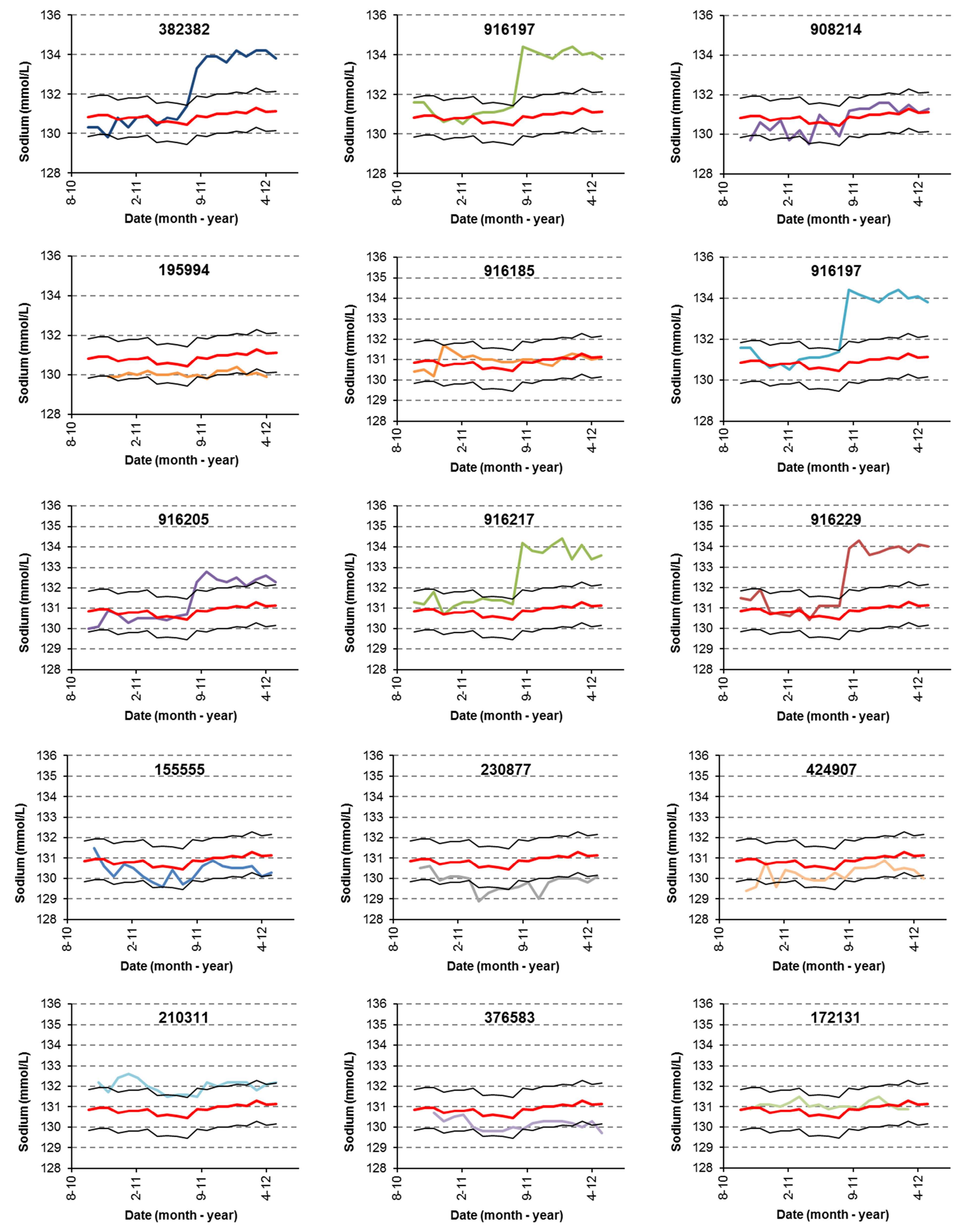
Supplemental Figure 6. Calcium IQC data from Biorad for several laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 0.05 mmol/L) in black lines.

Biorad IQC results for several laboratories on inorganic phosphor:



Supplemental Figure 7. Inorganic phosphor IQC data from Biorad. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 0.04 mmol/L) in black lines.

Biorad IQC results for several laboratories on sodium:



Supplemental Figure 8. Sodium IQC data from Biorad for several laboratories. Colored lines (not red) represent the data of the individual laboratory. The peer group data are shown with a red line, the applied QC limits (± 1 mmol/L) in black lines.

2 Chapter II

The Percentiler

This chapter is based upon:

- Goossens K, Van Uytfanghe K, Twomey P, Thienpont L, and Participating laboratories. Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make "Big Data" work. Clin Chim Acta 2015;445:12-18.
- De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru M, Thienpont LM, for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta 2016; doi: 10.1016/j.cca.2016.04.032. [Epub ahead of print]

Highlights

Most real-time laboratory quality monitoring is based on the analysis of QC materials. Although this is the most suitable tool for the first line of defense against measurement errors, it can miss some analytical variation in results from patient samples due to commutability issues of the samples. Therefore, as a surplus to the established IQC in detecting analytical errors, we will now discuss the developed Percentiler tool, which has the potential to visualize the analytical variation of performance on patient samples directly.

Highlights of the research:

- An automatic web application for monitoring patient medians ("The Percentiler") was built.
- The Percentiler does not require additional measurement, as is the case in traditional EQA, and makes use of perfectly "commutable" samples.
- The Percentiler data provide global evidence about test stability.
- Both individual laboratories and manufacturers can use the Percentiler to assess their problem analytes.
- The Percentiler is capable of showing what stability can be achieved by current state-of-the-art performance.
- The Percentiler can guide on which quality specifications, derived from state-of-the-art performance, to use.
- The Percentiler is capable of providing reliable peer group overviews to assess the relative comparability of the different assays on the market.

1. Introduction

As mentioned at the end of the previous chapter, most real-time laboratory quality monitoring is based on the analysis of materials which act as surrogate for a patient specimen (1). IQC, however, should not be used as a stand-alone tool to identify analytical errors. A second tool can be built on the direct use of results for patient sample analysis. This can be achieved by using procedures to detect absurd patient results and delta checks. While these use individual patient data, data can also be grouped as is done with the "average of normals" procedure (2,3). This concept was already introduced by Hoffman *et al* in 1965. Its use has been investigated in several settings, however, the basic concept was always the same: the average of patient results is calculated (e.g. daily or monthly), plotted in time, and compared with control limits (4-7). In the following chapters we will discuss two new applications which also make use of grouped patient data for the assessment of mid- to long-term quality monitoring: the Percentiler (chapter 2) makes use of daily patient medians and the Flagger (chapter 3) uses daily hypo- and hyper-flagging rates. These tools have the potential of directly showing the effect of analytical quality on results for patient sample analysis.

Before the Percentiler was created, a pilot study was conducted to verify whether it was possible to perform stability monitoring by plotting daily patient medians. When the patient medians for sodium from two laboratories were plotted, Stepman *et al* could easily follow the stability of the assays, and observe periods with a good stability versus periods where analytical shifts occurred (8,9). After this pilot study, the first foundation for the Percentiler project was laid in the thesis work of Dr. H. Stepman (10). The current study extended Stepman's work to 22 analytes and as much participants as possible.

The Percentiler is a freely available online web tool. The underlying database is fed with instrument-specific, daily patient medians. The latter are calculated and transferred by a laboratory from their middleware or Laboratory Information Software (LIS) to our database. This is done by e-mail. Subsequently, the data are automatically read by software into a MySQL database. In the pilot phase the project focuses on 20 clinical chemistry analytes, i.e., albumin (ALB); alkaline phosphatase (ALP); alanine aminotransferease (ALT); aspartate aminotransferase (AST); total-bilirubin (BIL); calcium (CA); total-cholesterol (CHOL); chloride (Cl); C-reactive protein (CRP); gamma-glutamyl transferase (GGT); glucose (GLU); potassium (K); creatinine (CRE); lactate dehydrogenase (LDH); magnesium (Mg); sodium (Na); inorganic phosphor (PHOS); total-protein (PROT); urea or bound urea nitrogen (BUN); and uric acid (UAC). In a later stage, free thyroxine (FT4) and thyroid-stimulating hormone (TSH) were added to the database, in the context of a project for standardization/harmonization of thyroid function tests.

2. Percentiler functionality

Patient medians are collected in a MySQL database and data are graphically and statistically interpreted via an online quality tool called “The Percentiler”. This site, which was created by IT engineer Bruno Neckebroek, is accessible to both the UGent project team and the individual participants. The latter group is only capable of interpreting their own data via the Quality and Samples tab; only the project team has the possibility to access all data and to create accounts for new laboratories.

Login

The Percentiler is accessible online on: https://www.thepercentiler.be/. Usernames (not case sensitive) and passwords (case sensitive) are available for individual labs, for groups (e.g. organizations with multiple sites), and for the entire database (see Figure 1). After sending data for eight weeks, the participant receives its login information, which gives access to the user interface for the graphical interpretation of his/her data. Those who do not yet participate can make use of a demo-account (username: DEMOLAB; Password: demo1234). Note that for the demonstration laboratory data of only one year are used. User guides are available in Dutch, English and French so that interested laboratories and new participants can get acquainted with the tool.

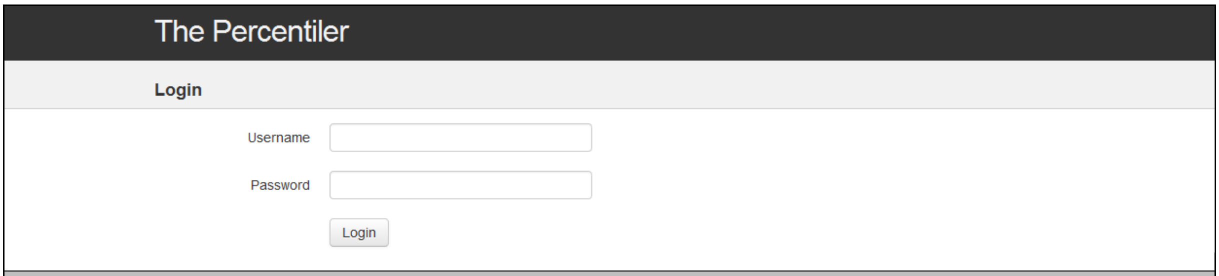
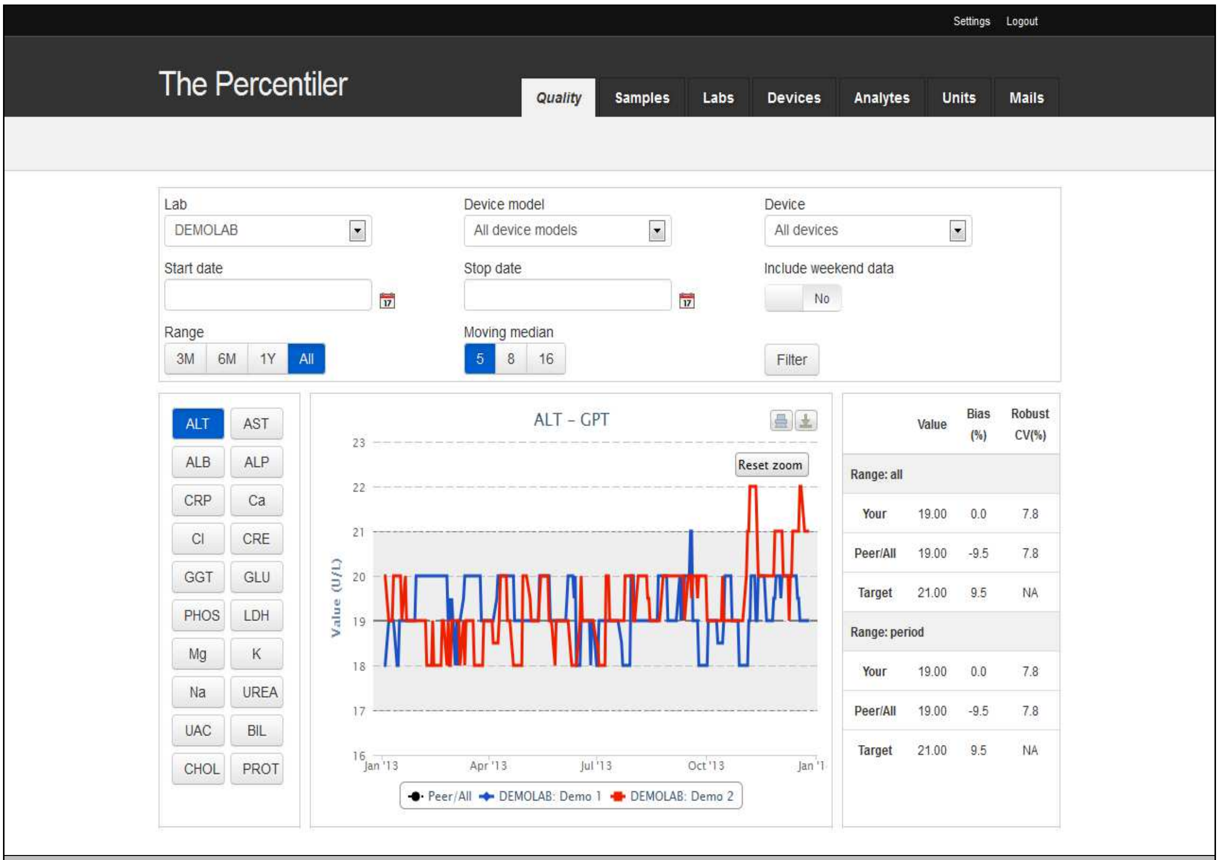


Figure 1. The Percentiler login screen.

The “Quality” tab



Lab: DEMOLAB

Device model: All device models

Device: All devices

Start date:

Stop date:

Include weekend data: No

Range: 3M 6M 1Y [✓ All]

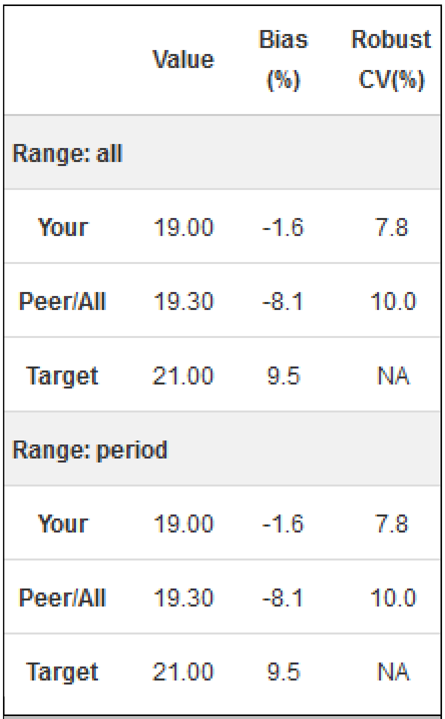
Moving median: [✓ 5] 8 16

	Value	Bias (%)	Robust CV(%)
Range: all			
Your	19.00	0.0	7.8
Peer/All	19.00	-9.5	7.8
Target	21.00	9.5	NA
Range: period			
Your	19.00	0.0	7.8
Peer/All	19.00	-9.5	7.8
Target	21.00	9.5	NA

Figure 2. Demonstration of “The Percentiler”, enabling a laboratory to track its moving medians per instrument over time (Quality tab). The chart shows for the selected lab i) the moving median of the selected devices over the selected period (full colored lines); ii) the long-term median (horizontal long-broken grey line; represents the median calculated from all daily patient medians provided by the laboratory to which the graph applies) and the limits for the respective analyte (short-broken grey lines; the area in-between is shaded); iii) the Peer Group moving median (black broken line).

This window shows the stability information from a selected laboratory in a graphical and tabular output. The lab can be chosen in the “Lab” box. Only the project team has access to the data of all laboratories, while participants are restricted to view only their own data. In the “Device model” box one can choose either “your device” or “All device models”. in the “Device” box, one can choose between the different devices of the chosen lab. In the default condition all devices of the lab are shown. In addition, it is possible to include or exclude weekend data, and to choose the data range via “Start date”, “Stop date” or via the buttons 3M (months), 6M, 1Y (year), and All. Finally, one needs to select *n* for the moving median (default: 5) and the analyte with the respective buttons. Once selections are done, the chart appears and the table is filled. The chart (example in Figure 2) shows for the selected lab i) the moving median of the selected devices over the selected period (full colored lines); ii) the long-term median (long-broken grey line; represents the median calculated from all daily medians from results from patients provided by the laboratory to which the graph applies) and the limits for the respective analyte (short-broken grey lines; the area in-between is shaded and called the stability zone); and iii) the peer group moving median (black broken line). It is possible to print and download the chart by pressing the respective buttons.

The table from the Quality tab (example in Figure 3) shows two parts: the upper part refers to the data of the complete period (static when the range is changed), the lower part refers to data of the selected range (dynamic when the range is changed).



	Value	Bias (%)	Robust CV(%)
Range: all			
Your	19.00	-1.6	7.8
Peer/All	19.30	-8.1	10.0
Target	21.00	9.5	NA
Range: period			
Your	19.00	-1.6	7.8
Peer/All	19.30	-8.1	10.0
Target	21.00	9.5	NA

Figure 3. Demonstration of “The Percentiler” (Quality tab) – Table with summary statistics (bias, robust CV)

- The “Value” column lists:
  - “Your” median = the long-term median from the selected laboratory/instrument. This corresponds to the horizontal long-broken grey line in the Percentiler graph.
  - “Peer/All” median = the long-term median from either the peer group data or all data.
  - Target = a fixed “reference” value which is taken from literature or trueness-based reference intervals (see pages 69-70).

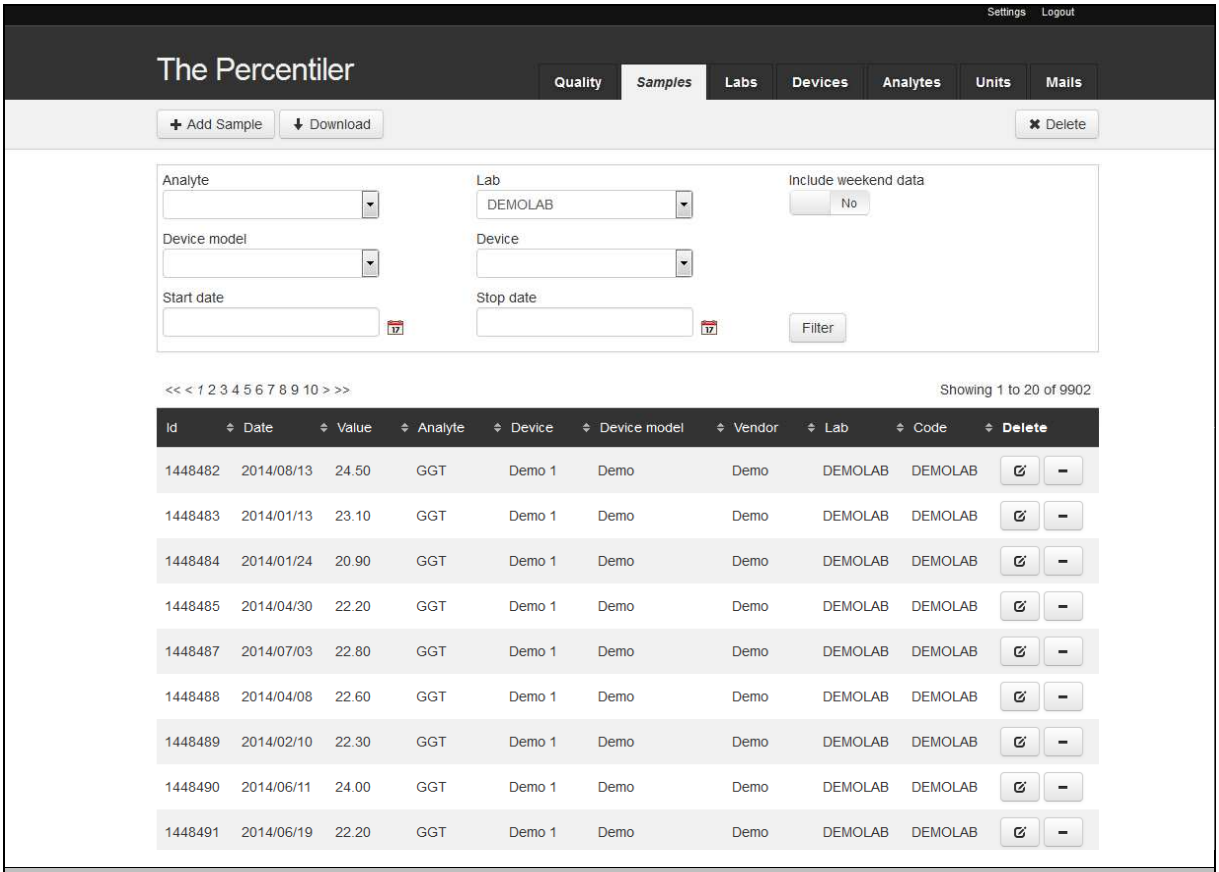
- The “Bias (%)” column indicates:
  - “Your” bias = bias of the selected laboratory/instrument compared to the peer group.
  - “Peer/All” bias = bias of the peer/all data compared to the target.
  - Percentiler limit expressed as a percentage of the target (e.g. 9.5% = 2 U:L of 21 U/L).

- The “Robust CV (%)”, calculated as 100 * (median absolute difference x 1.4826)/(median), lists the lab’s robust CV (“Your”), and the robust CV of its Peer/All.

The robust CV relates to the regular CV as the median relates to the mean. When data is perfectly normally distributed the median will equal the mean, and the robust CV will equal the regular CV. When data isn’t normally distributed the robust CV is less sensitive towards outliers, because it uses the median absolute difference as robust measure of the variability of quantitative data. This median absolute difference is calculated by taking the median of the absolute differences of each measurement in the dataset with the median of that dataset.

The “Samples” tab

This window shows all entries in the database of the chosen laboratory (example see Figure 4). Note, however, that data are mapped to the “Percentiler” analyte names and units (not necessarily identical to the names and units given by the participants). One can filter the data (for example, according to analyte, to select in the “Analyte” box, always followed by pressing the “Filter” button) and sort (▲▼) (e.g. according to date, value, etc.). This tab also contains an Excel exporting feature, where one can download the data, either completely or filtered. The project team uses this tool to make monthly overviews of the data per analyte and per peer group. These overviews can be used to visualize potential changes in peer group medians (which is discussed on pages 81-82).



Analyte:

Lab: DEMOLAB

Include weekend data: No

Device model:

Device:

Start date:

Stop date:

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Showing 1 to 20 of 9902

Id	Date	Value	Analyte	Device	Device model	Vendor	Lab	Code	Delete
1448482	2014/08/13	24.50	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448483	2014/01/13	23.10	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448484	2014/01/24	20.90	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448485	2014/04/30	22.20	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448487	2014/07/03	22.80	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448488	2014/04/08	22.60	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448489	2014/02/10	22.30	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448490	2014/06/11	24.00	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	
1448491	2014/06/19	22.20	GGT	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	

Figure 4. Demonstration of “The Percentiler” (Samples tab). The table shows nine attributes for the selected lab: lab identification number; date; median value; analyte; device identification; vendor or manufacturer; laboratory name; and lab ID code.

The “Labs” tab and data transmission

Via the “Labs” tab the project team can map new participants. Each laboratory is given a unique identifier and code, usually chosen by the participants themselves. Each data format provided by the laboratory (date, device representation, analyte coding, and unit coding) is mapped according to the Percentiler format.

The data of each participant should be in a “table” consisting of a data-related number of rows with ten fields (note, fields 8 to 10 are only necessary for mapping in the Flagger, see chapter 3). The order of the fields can vary.

1. Laboratory ID: chosen by the laboratory or already assigned by the laboratory
2. Date
3. Instrument ID: as already assigned by the laboratory
4. Patient code: three letters (e.g. OUT) or as already assigned by the laboratory
5. Analyte name: as already assigned by the laboratory
6. Unit: as already assigned by the laboratory
7. Value of the median (decimal separator = point or comma)
8. Number of results used for calculating the median
9. % flagged hypo
10. % flagged hyper

Data can be transmitted as: i) e-mail embedded table; ii) e-mail attached EXCEL-file; or iii) e-mail attached text-file. Data are sent to a database-specific email address (percentile@stt-consulting.com) and is automatically read into a MySQL database. Examples of data transition are shown in Figure 5.

e-mail embedded Table (without *n* and flagging frequencies!)

```
From: ***
Sent: Saturday, 28 september 2013 06:31
To: percentile@stt-consulting.com
Content: Empower Percentile Project

Time produced : 27-09-2013 00:00 - 27-09-2013 23:59

ABCDEF;27/09/2013;C16000-5;POL;NA;mmol/L;140.9
ABCDEF;27/09/2013;C16000-6;POL;NA;mmol/L;139.4
ABCDEF;27/09/2013;C16000-5;POL;CL;mmol/L;104.6
ABCDEF;27/09/2013;C16000-6;POL;CL;mmol/L;103.8
ABCDEF;27/09/2013;C16000-5;POL;CA;mmol/L;2.44
ABCDEF;27/09/2013;C16000-6;POL;CA;mmol/L;2.42
ABCDEF;27/09/2013;C16000-5;POL;P;mmol/L;1.09
ABCDEF;27/09/2013;C16000-6;POL;P;mmol/L;1.03
ABCDEF;27/09/2013;C16000-5;POL;MG;mmol/L;0.85
ABCDEF;27/09/2013;C16000-6;POL;MG;mmol/L;0.84
ABCDEF;27/09/2013;C16000-1;POL;UREUM;mg/dl;30.5
ABCDEF;27/09/2013;C16000-2;POL;UREUM;mg/dl;33.3
```

Text attachment to e-mail

```
From: ***
Sent: Saturday, 28 september 2013 06:31
To: percentile@stt-consulting.com
Content: Empower Percentile Project

Time produced : 27-09-2013 00:00 - 27-09-2013 23:59

Filename: Empower Percentile.txt
ABCDEF;27/09/2013;80_AU5822;POL;ALB;g/dl;3.0;69;5;2
ABCDEF;27/09/2013;80_AU5822;POL;CA;mmol/l;2.125;90;10;12
```

EXCEL attachment to e-mail (without n and flagging frequencies!)

```
From: ***
Sent: Wednesday, 13 november 2013 06:18
To: percentile@stt-consulting.com
Content: Empower Percentile Project

Time produced : 12-11-2013 00:00 - 12-11-2013 23:59

Filename: Empower Percentile.xlsx (or xls)
```

ABCDEF	12/11/2013	VITROS5.1FS	E	Alb	g/L	42.2
ABCDEF	12/11/2013	VITROS5.1FS	E	APase	U/L	91.5
ABCDEF	12/11/2013	VITROS5.1FS	E	Ca	mmol/L	2.355

Figure 5. Examples of data transition to the MySQL database of the Percentiler.

The “Devices” tab

In the “Devices” tab the project team can define new device vendors (or manufacturers), device models (or peer groups), and device types. In the device model category, it is possible to define a new peer group and set a starting date for showing the peer target on the chart. When a new device is mapped for a participant, it is catalogued with the correct vendor, model, and type, thereby allowing comparison of its data to the correct peer group.

The “Analytes” and “Units” tab

In the “Analytes” tab the project team can define new analytes. Each analyte is given an ID, a unit, a fixed target value and an allowable range for stability. Units are chosen according to the international system (SI) of units. When a laboratory reports non-SI units these are converted to the SI target by providing a factor. In the “Units” tab every unit which is used for reporting is defined.

The “Mails” tab

When a laboratory sends its data, this is registered in the “Mails” tab. The table in this sheet shows: the e-mail sender, date and time of receipt, e-mail subject, the number of median values reported, and when relevant an error message. For example, when a laboratory reports data for a new device which is not yet mapped in the Percentiler, an error message mentions the reporting of an unknown device.

3. Data stratification for outpatients

Preferably, laboratories should only send daily medians on their outpatients, i.e., patients who are not hospitalized overnight but who visit a hospital, clinic, or associated facility for diagnosis or treatment. Clinical laboratories, particularly hospital laboratories, however, receive samples from many different sites:

- The hospital itself (inpatients, i.e. patients who stay for one or more nights in a hospital for treatment, often specified by ward)
- Special treatment units (diabetic centers; dialysis centers; wellness centers; nursing homes)
- General practitioners/specialists who send samples from chronically diseased patients
- Samples from "New" patients sent by General Practitioners
- Samples from patients who have a general health check (also sent by General Practitioners)

The middleware/LIS should be able to code these different "sender-sites" by type (the actually coded types will depend on the situation in the laboratory). "Outpatients" may then be defined by the laboratory in principally two different ways:

- "Total minus" (total minus inpatients, minus dialysis, for example)
- "Add-on": samples from general practitioners + chronically diseased + wellness + others.

The driver should support the "Total-minus" solution; this allows the laboratory a step-wise stratification of their patient population. In addition, the middleware/LIS should be able to code the results by measuring instrument; for example, when there are two Cobas c701 instruments on a "Cobas 8000 analyzer" those should be given two different identifiers (for example: Cobas c701 A, Cobas c701 B).

Comparison of outpatient/all patient monitoring

Table 1 shows the differences between all patient and outpatient medians in typical hospital laboratories and the CV-ratio (not the robust CV) of those two groups. The patient medians and CV-values from both outpatients and all patients were calculated from the same dataset comprising two months of measurements from one instrument from a single medium sized laboratory. Note, although data were extracted for a single hospital laboratory, they were confirmed with data from other laboratories. There is a considerable difference in the median of outpatients and all patients for most of the analytes. Note, however, only outpatients will give medians which can be compared with "target" medians from reference interval information, because the latter information is based on the distribution of a “healthy” population.

The CV-ratios All-/Out-patients are typically <1, meaning that monitoring all patients would result in better stability. We decided, however, to continue with outpatient monitoring because only outpatients may give comparable values between different laboratories, allowing the assessment of laboratory/peer group bias. The somewhat higher variability can be compensated by a slightly higher “n” for the moving medians. In cases where an exceptionally high population variation was observed for a laboratory, we proposed them to define two Percentiler accounts. Data from outpatients could still be used for comparison with target medians, whilst the second account with all data was more suited for stability monitoring.

Table 1. Comparison between outpatient and all patient data.

	%-Diff (Out/All)*	CV Ratio (All/Out)
Albumin	14.2	1.76
ALP	-7.0	0.69
ALT	-4.6	0.79
AST	-4.5	0.83
Total Bilirubin	-7.3	0.68
Calcium	3.0	1.18
Total cholesterol	1.6	0.90
Chloride	0.2	0.76
Creatinine	1.9	0.74
CRP	-93.2	1.27
GGT	-23.6	0.89
Glucose	-5.2	0.87
Potassium	2.8	0.58
LDH	-5.0	0.83
Magnesium	1.6	0.61
Sodium	0.4	0.89
Inorganic phosphorus	0.7	0.65
Total protein	7.0	1.13
Urea	3.4	0.65
Uric acid	4.4	0.78

$$*\%\text{-Diff} = \frac{\text{Outpatient median - All patient median}}{\frac{1}{2}(\text{Outpatient median + All patient median})} \times 100$$

4. Limits and targets

Stability limits

The Percentiler bias limits (or specifications), indicated with the horizontal, grey broken lines in the graphs, are guided by the systematic error goals based on biological variation ("desirable" values from the Westgard website; 11). However, we took the current capability of diagnostic assays into account (i.e. state-of-the-art performance) and expanded these goals for analytes with narrow biological variation, such as sodium, chloride, calcium, etc. In contrast, we restricted the upper limit to ~10% for analytes with very high biological variation (e.g. CRP). The values for each limit are mentioned in Table 2 of Article 1. They are assigned semi-arbitrarily for each analyte (by spending a lot of time interpreting the Percentiler data) so most of the participating laboratories belonging to the best performing peer group are able of achieving the set limits. The actually chosen numbers were "tailored" to the SI-units. For albumin and total-protein, for example, we chose 1 g/L; the respective percentages were then calculated at the median concentration, resulting in "non-integer" numbers: 2.4% for albumin, 1.4% for total-protein.

Note, the Percentiler limits are dynamic, meaning they are adapted according to changes in state-of-the-art performance and changes in the biological variation data in the Westgard Biodatabase. Therefore, some Percentiler figures in this thesis might still show a grey stability zone flanked by limits representing former numbers. Whenever the limits are functional to the provided explanation, however, the values and figures for the limits are adapted to the most recent numbers.

Target values

Target values are structured according to hierarchy: long-term laboratory median, peer group moving median, and a "reference" target. The latter, however, is difficult to define. The Nordic Reference Interval Project (NORIP) database is the only source we know which claims "trueness-based" reference intervals (12). The reliability of that source is high for analytes such as sodium and calcium. The information for some enzymes has to be used critically (for example, there have been changes in the procedures recommended by the International Federation of Clinical Chemistry and Laboratory Medicine or IFCC).

We also compiled reference interval information from manufacturers' data sheets. For example, for albumin we compiled the following reference intervals: Abbott Architect: [35-50 g/L]; Beckman AU: [35-52 g/L]; Beckman DxC: [35-50 g/L]; Ortho Vitros: [35-50 g/L]; Roche Cobas: [35-52 g/L]; Roche Integra: [35-52 g/L]; Siemens Advia: [32-48 g/L]; and Siemens Vista: [35-52 g/L]. The mean of each interval was calculated and converted to a median value using the mean-median relationship from the NORIP database. For albumin those median values were: 42.3; 43.3; 42.3; 42.3; 43.3; 43.3; 39.8; and 43.3 g/L respectively. The median value of those assay medians was calculated (i.e., 42.8 g/L for albumin) and cross compared with the NORIP data. This comparison may help to define some preliminary “reference” targets for several analytes (see Table 2). Current analytes with “reference” targets could be albumin, ALT, AST, calcium, glucose, magnesium, sodium, phosphate, urea, uric acid.

Table 2. Compiled reference interval information from manufacturers’ data sheets. The median value from the reference intervals is compared with the NORIP data.

		Median		**Diff**
	Unit	**NORIP**	**Assays**	**(%)**
Albumin	g/L	41.5	42.8	**3.1**
ALP	U/L	63.0	73.2	16.2
ALT	U/L	21.0	21.5	**2.3**
AST	U/L	23.0	22.6	**-1.9**
Total Bilirubin	µmol/L	10.0	8.6	-13.8
Calcium	mmol/L	2.34	2.33	**-0.3**
Total cholesterol	mmol/L	5.20	-	-
Chloride	mmol/L	-	*102.5*	-
CRP	mg/L	-	-	-
GGT	U/L	22.0	17.5	-20.6
Glucose	mmol/L	4.87	4.80	**-1.5**
Potassium	mmol/L	4.05	4.25	4.9
Creatinine	µmol/L	70.7	77.0	8.9
LDH	U/L	152	176	15.6
Magnesium	mmol/L	0.83	0.86	**3.6**
Sodium	mmol/L	141.1	140.6	**-0.4**
Inorganic phosphorus	mmol/L	1.13	1.12	**-0.9**
Total protein	g/L	69.8	73.2	4.9
Urea	mmol/L	4.89	4.96	**1.4**
Uric acid	µmol/L	282	272	**-3.5**

5. Comparison with IQC

Table 3. Comparison of typical CV-values for IQC-monitoring and percentile monitoring.

	IQC - CV (%)	Percentile - CV (%)	CV-ratio
Albumin	0.9	2.0	2.3
ALP	1.1	5.4	4.8
ALT	2.8	5.8	2.1
AST	1.9	4.7	2.4
Total Bilirubin	2.2	12.1	5.6
Calcium	1.0	1.2	1.2
Total cholesterol	1.5	5.4	3.6
Chloride	0.4	0.6	1.2
CRP	2.6	23.5	8.9
GGT	1.8	11.1	6.2
Glucose	0.6	2.8	4.4
Potassium	0.4	1.4	3.7
Creatinine	1.6	4.7	3.0
LDH	1.3	4.9	3.8
Magnesium	0.9	2.6	3.1
Sodium	0.3	0.4	1.4
Inorganic phosphorus	1.0	3.7	3.8
Total protein	0.9	1.9	2.1
Urea	1.2	4.8	4.0
Uric acid	0.5	5.3	10.5

Table 3 compares typical CV-values for IQC-monitoring and percentile monitoring (daily medians) for UZ Gent. These CV -values (not the robust CV-values) were calculated from a single dataset comprising 40 days of measurements from one instrument using a single reagent lot. The CV's for patient data are, generally, greater than for IQC data (ratios >1). Exceptions are the "high-volume" analytes with narrow biological variation (calcium, chloride and sodium). Therefore, daily QC decisions will have to rely on IQC data in most laboratories. However, applying moving medians with "tailored" n, will reduce the variability by $\sqrt{n}$. This makes percentile monitoring applicable for mid-to long-term monitoring of analytical variability. The actual number of daily medians required for reliable monitoring will depend on the number of outpatient results generated by a laboratory, in combination with the population variation of a certain analyte (again, analytes with low biological variation allow lower “n” to be used).

By way of illustration, we compared daily patient medians with IQC results from that same dataset. Figure 6 shows the moving averages grouped per four days (not the moving medians in this case!) for daily patient medians in blue and daily IQC results in black for four analytes. Note, IQC values were normalized towards the patient medians. Analytes with CV’s for patient data which are similar to IQC, tend to follow the same pattern for both types of data (e.g. calcium and chloride). Analytes with higher CV’s based on patient data can only be used for mid- to long-term monitoring of the patient medians (e.g. glucose and uric acid). An overview of all analytes can be found in the annex (pages 104-106).

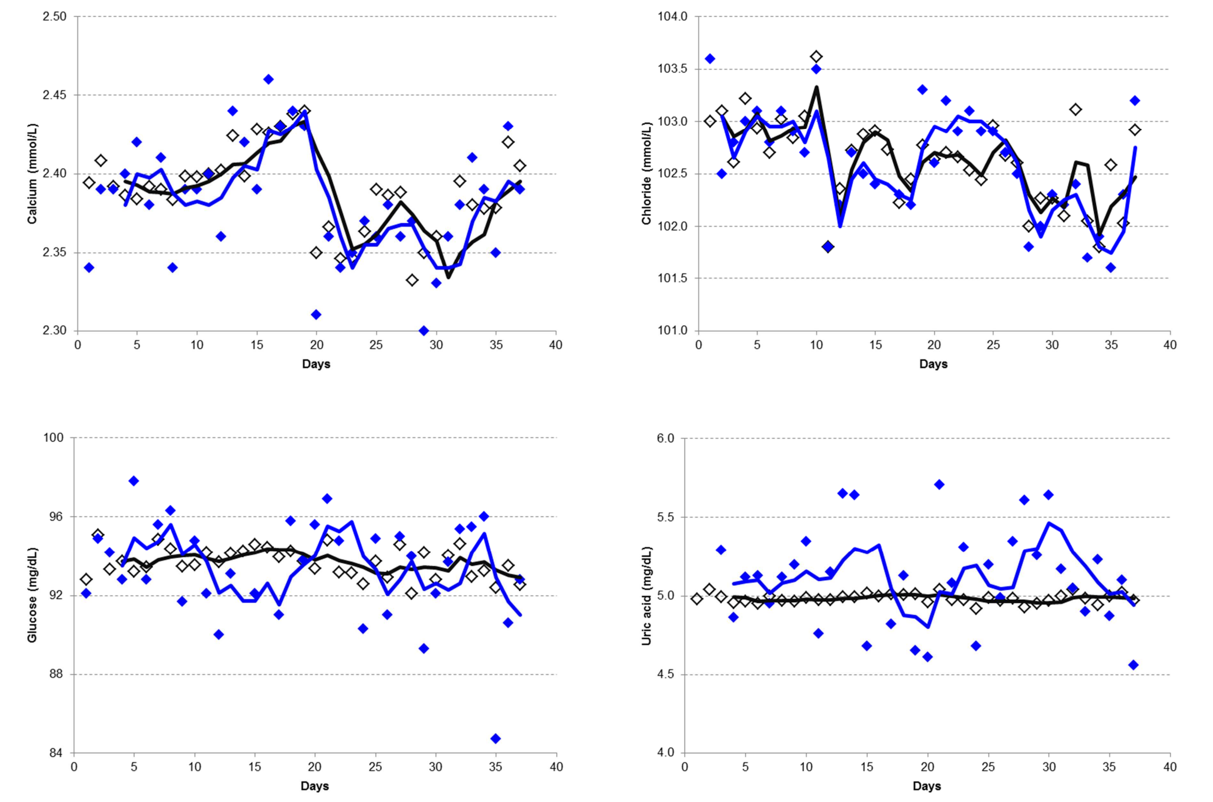


Figure 6. Comparison of moving averages (n = 4 days) for daily IQC values (black) and daily patient medians (blue) for four analytes (calcium, chloride, glucose and uric acid). Moving averages were calculated from a single dataset comprising 40 days of measurements from one instrument using a single reagent lot.

6. Percentiler applications

Mid- to long-term stability assessment

A) Big private laboratories versus small hospital laboratories

All sorts of clinical laboratories can participate to the Percentiler. The size of these laboratories will, however, influence how they need to interpret the graphical output from the Percentiler and what type of information they can extract from it. Laboratories of all sizes participate ranging from small (~200 beds) and big hospital laboratories (~2000 beds) to high-throughput private laboratories. For smaller hospital laboratories, with greater intrinsic median variation, a higher number of days (n = 8 or 16) has to be chosen for the moving median, especially for analytes with big population variation such as CRP, GGT, albumin, and uric acid. Consequently, the effect of that population variation is partially suppressed, but the consequence is that longer observation times are needed to uncover analytical instabilities and that some shorter-term information is lost (see Figure 7). High-throughput laboratories will be able to work with a moving median grouped per 5 days and they might even be capable of observing small analytical effects which are smaller than the Percentiler limits.

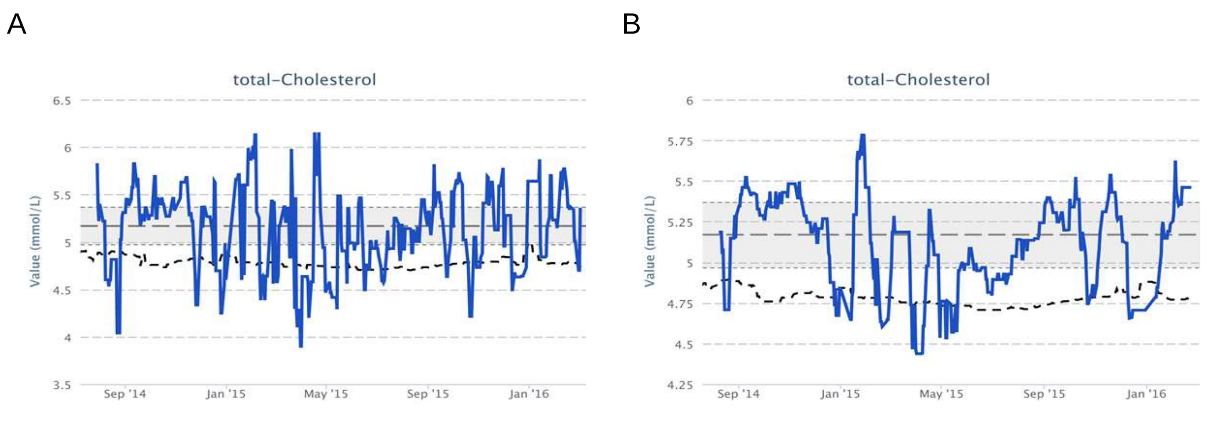


Figure 7. Higher population variation in hospital/low-throughput laboratories can be partly compensated for by increased “n” for calculation of the moving median. (A) n = 5 days. (B) n = 16 days.

Laboratories with high population variation, however, might also run in danger of overinterpreting the Percentiler graphs. Interpreting the stability for total-cholesterol, in Figure 7, when the moving median is grouped per 16 days (right graph), might suggest the occurrence of high instability. These observations are not or barely confirmed when the moving median is grouped per 5 days (left graph). Interpretation of the Percentiler data when the population variation is high should always be handled with caution. Therefore, laboratories with high population variation are advised to verify their stability on a moving median of n = 5 (or n = 8), when they are interpreting their long-term stability at n = 16.

B) Stability issues in the laboratory or in the entire peer group

The Percentiler provides a real-time monitoring tool for test stability across laboratories and manufacturers. The individual laboratory can use it as a direct, real-time quality indicator for its analyses of patient samples, not confounded by commutability issues. Together with data from IQC, the project helps the laboratory to build evidence about mid- to long-term stability of assays/performance. A number of examples where laboratories can observe significant analytical variation are shown in Figure 8.

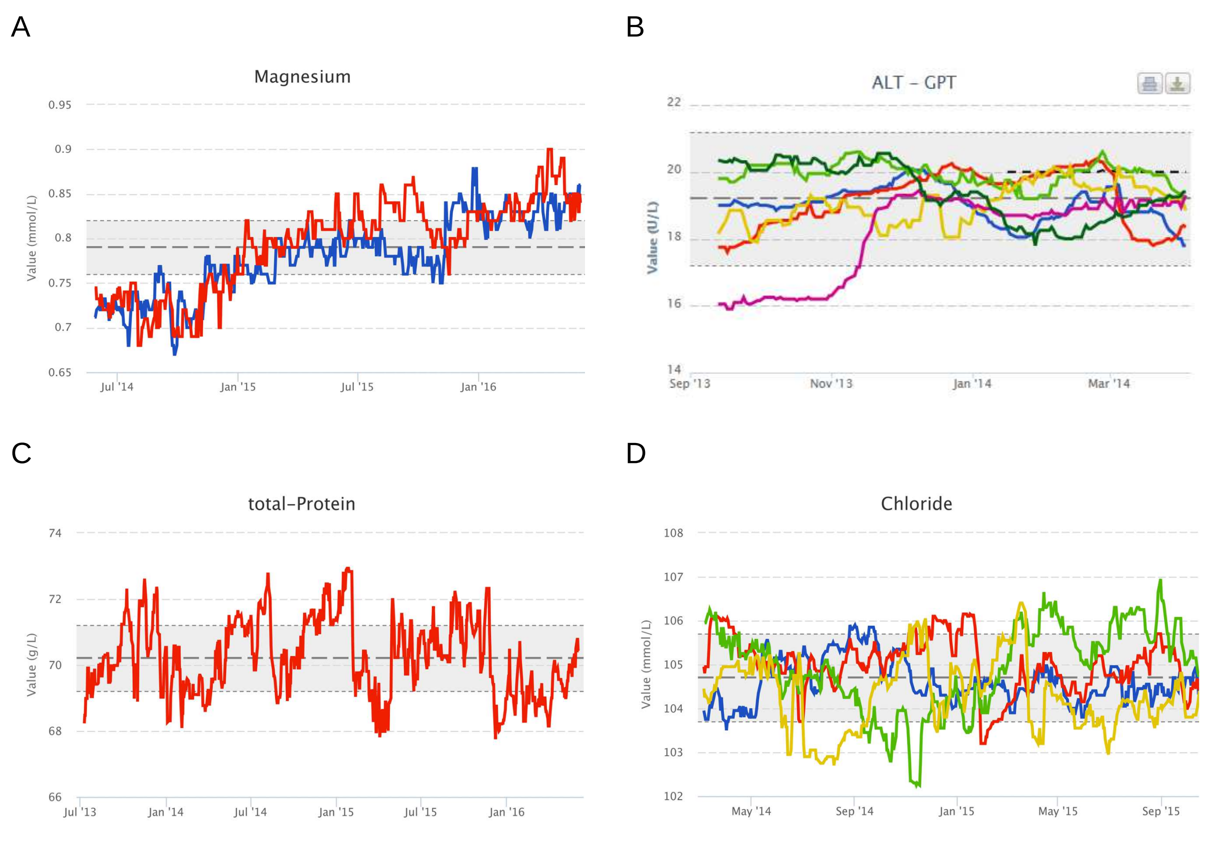


Figure 8. Selected time course examples from The Percentiler (stability limits in gray). (A) Long-term drift for a magnesium assay caused by several upwards lot shifts (B) Difference between instruments for an ALT assay: one system lower for long period. (C) Sawtooth pattern for a total-protein assay: instrument instability causes the instrument to drift which is, after a while, corrected by a recalibration (shift). (D) Several long-term drifts and shifts leading to non-comparability between four different chloride assays.

Of course the single laboratory can’t distinguish whether stability issues are related to its own performance or that of the manufacturer’s instrument. To know whether a problem is restricted to a specific laboratory, occurring in an entire peer group, or in the entire Percentiler database, one needs access to the entire database. That’s why the project team informs participants (both laboratories and manufacturers) about problems occurring in several labs belonging to the same peer group, by means of peer group specific reports. In addition, manufacturers are, thereby, provided with real-time data about field performance of their instruments and application issues in particular laboratories. Figure 9 shows a returning problem for the total-bilirubin assay in a single peer group. These types of observations point to the need for an improvement of the assay for an entire peer group.

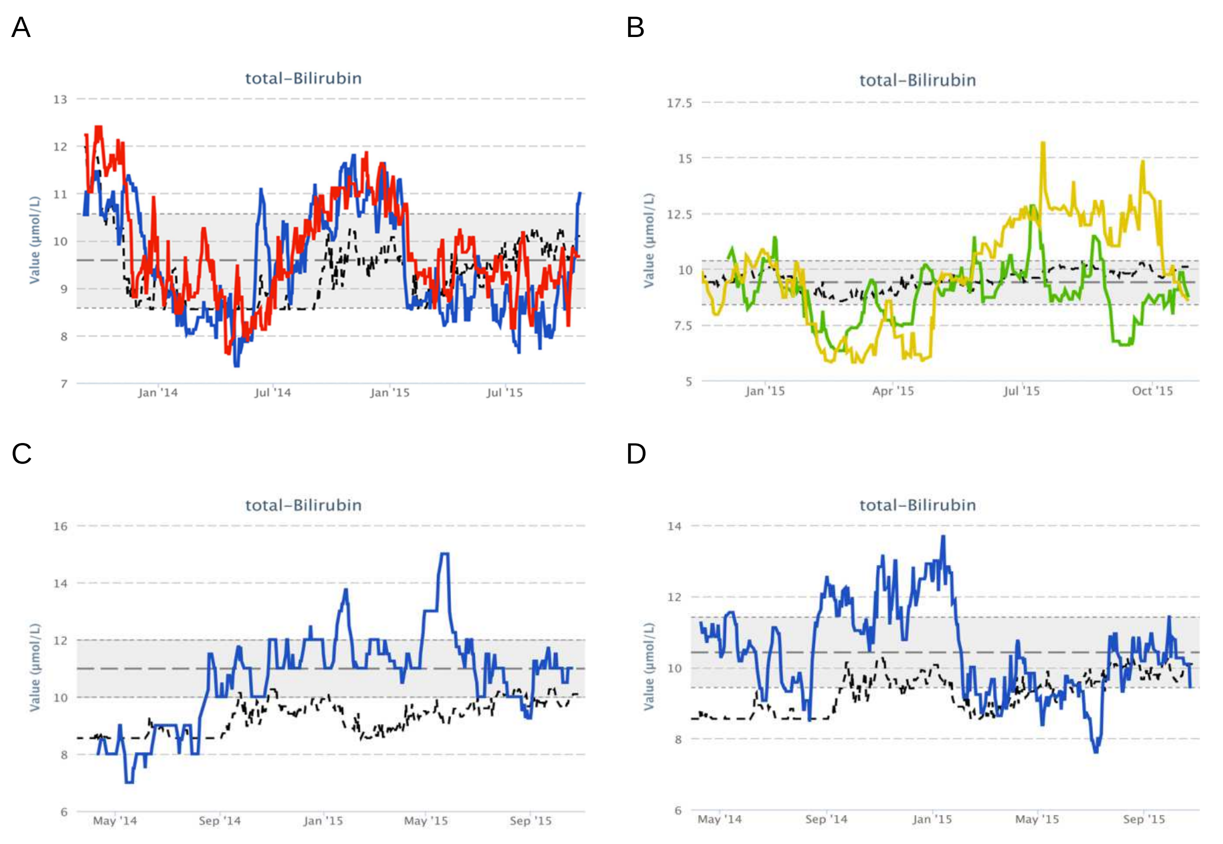


Figure 9. Selected time course examples from The Percentiler (stability limits in gray). Stability issues for the total-bilirubin assay in several laboratories belonging to the same peer group.

Assessing stability with the Percentiler mostly involves investigation of the graphs to detect significant analytical shifts and drifts. Visual inspection of all graphs, however, is a laborious work, and therefore it would be interesting to interpret stability in a more automated manner. One possibility would be to use the (robust) CV, mentioned in the descriptive table of the Percentiler application. Similar to using the CV as a measure for (in)stability with IQC, the robust CV from the Percentiler might be a good indicator to compare the stability between different periods for a single laboratory/instrument. The potential of using the robust CV for that matter, however, hasn't been thoroughly investigated. One might assume that this approach can work well for laboratories that have low population variation and a stable population (stratification) throughout the year. On the other hand, there is no guarantee that an increase in the robust CV might always be regarded as a deterioration of the stability for an instrument (especially for small sized labs). One would always need to exclude whether a population effect might cause the increase in the robust CV. In other words, investigating how the Percentiler data can be used to assess stability in an automated manner is still open for improvement.

C) Case study – instability of ALT assays jeopardizes patient management

The value of the Percentiler was considered in an ALT case study. Literature indicates the relevance of mild elevations of transaminases (ALT and AST) in nonalcoholic fatty liver disease (metabolic syndrome), which becomes more and more prevalent as obesity rates increase (13). Although the actual ALT and AST values may differ from laboratory to laboratory, serum levels are usually considered normal if they are less than 40 U/L for AST and less than 50 U/L for ALT. However, some experts have suggested lowering the upper limit of normal because of the increasing prevalence of obesity and associated nonalcoholic fatty liver disease, which may not be detected using the traditional, higher normal values (14). A case study (female developing metabolic syndrome) was considered with the following ALT values (U/L): 17, at the start of yearly check-up; 44, diagnosed as development of metabolic syndrome; 28, after six months of treatment; 18, after long-term treatment. This case study showed that, for correct monitoring of disease development and treatment, stable and comparable assays are needed. However, in the Percentiler application, significant lot-to-lot changes are sometimes observed for these assays (for example: ALT 35 to 22 U/L, in several steps; Figure 10A), which shows that these can jeopardize patient management.

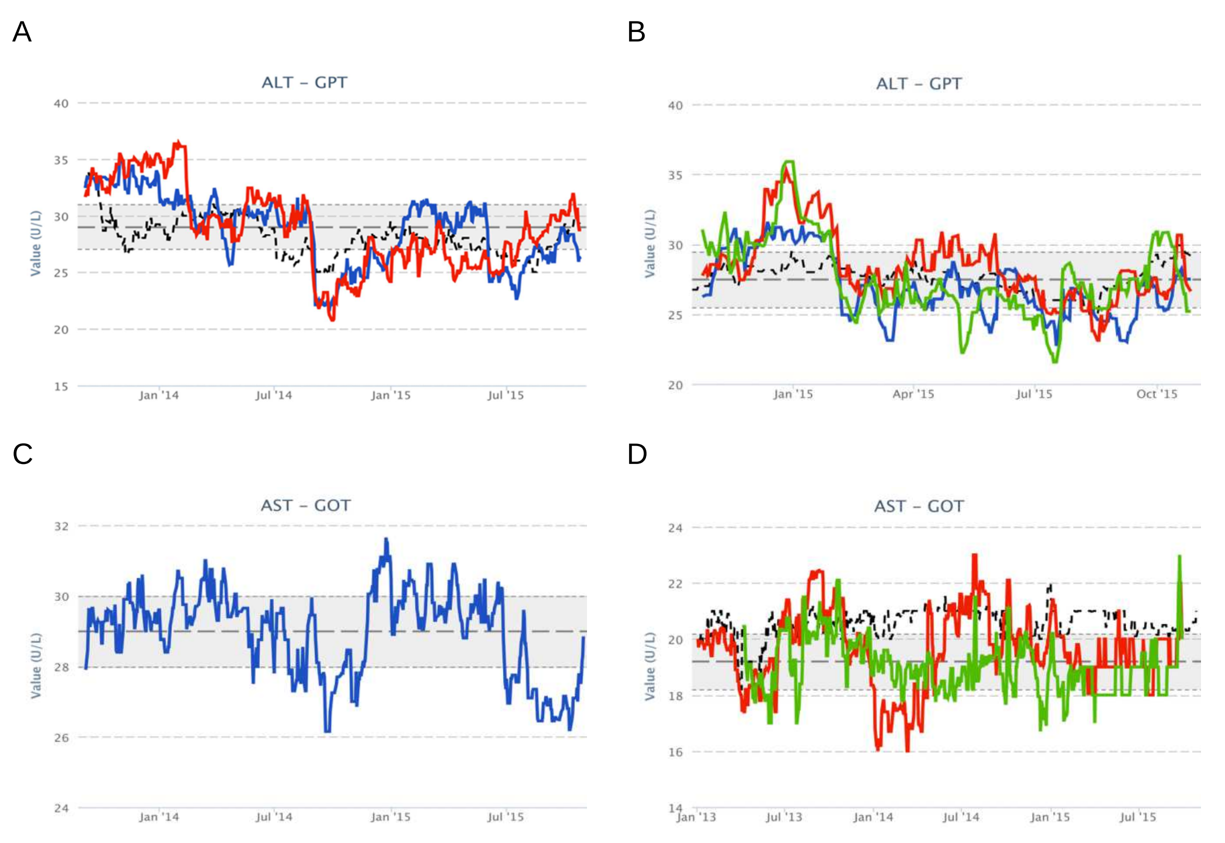


Figure 10. Selected time course examples from the Percentiler application indicating the instability of some ALT and AST assays.

D) Preanalytics

A typical observation when blood samples are stored or transported for a longer time, is the occurrence of a temperature-dependent pattern for certain analytes, e.g., potassium and LDH. At lower temperatures the erythrocyte/plasma potassium gradient can no longer be kept stable if the blood is not centrifuged on time (the responsible enzymes need a certain temperature for activity). As a result, the values of the patient potassium medians are elevated during the winter (15). For LDH the reverse pattern is observed, with the highest values in the summer. The latter can be explained by increased hemolysis of red blood cells at higher temperatures (16). These patterns are most often observed in private laboratories, when the clinical laboratory is at a different location than where the blood is drawn and not centrifuged (see examples in Figure 11).

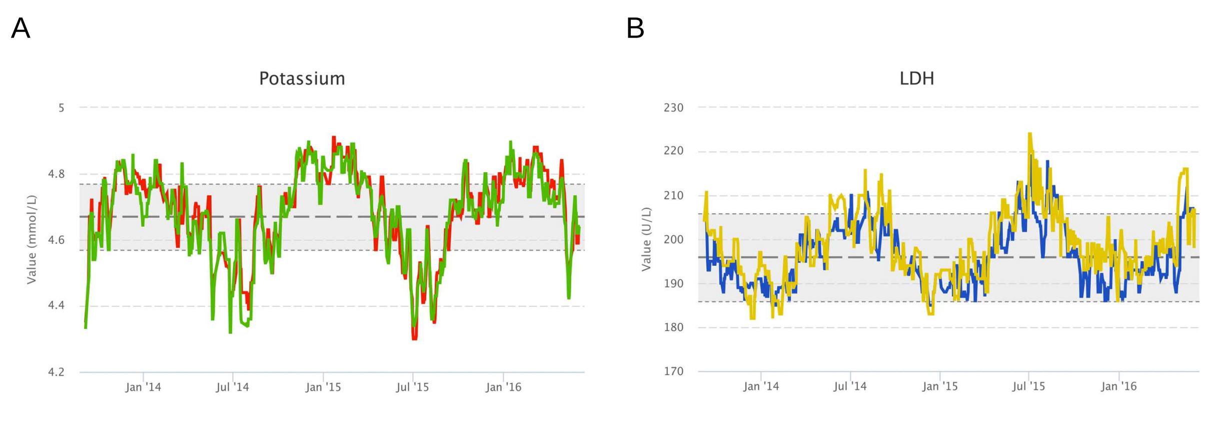


Figure 11. Selected time course examples from the Percentiler application indicating preanalytical variation for (A) potassium and (B) LDH. For both analytes a sinusoidal pattern occurs which repeats every year.

E) Seasonal variation

In some cases seasonal variation of the patient medians can be observed in the Percentiler for certain analytes. For example, several laboratories showed a seasonal variation in the moving medians for TSH (Figure 12), which is also reported in literature (17,18). However, in order to fully exploit the Percentiler to visualize seasonal variation, preferably data from big-sized laboratories with proven stable analytical performance and low population variation should be used. A more extensive discussion on the seasonal variation in the Percentiler can be found in an article from De Grande *et al* (19).

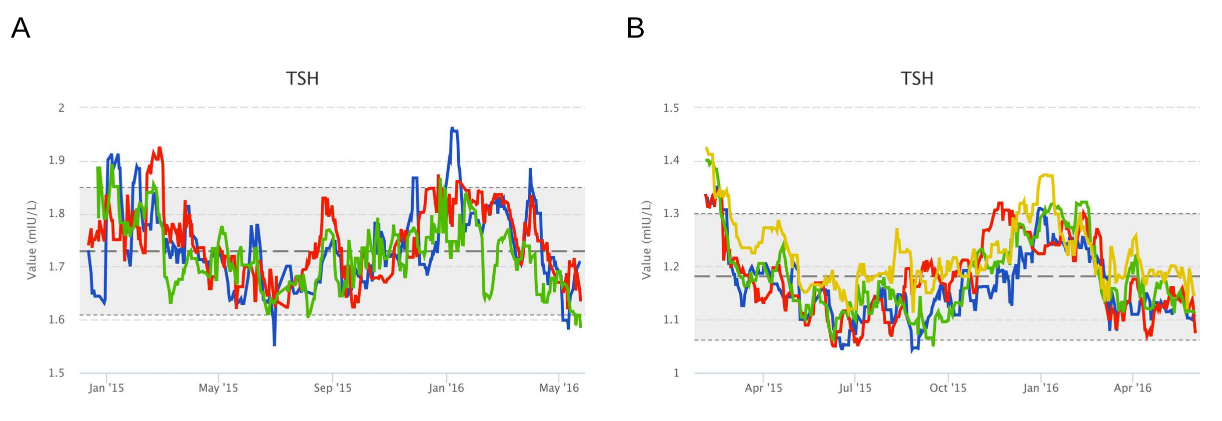


Figure 12. Selected time course examples from the Percentiler application showing seasonal variation in the moving medians for TSH.

Bias assessment

A) Bias assessment between different instruments from a single laboratory

The Percentiler not only provides a real-time monitoring tool for test stability, it can also be used for bias assessment. First of all, laboratories can verify whether their in-house instruments are comparable. If the long-term patient medians for their instruments differ, this can have two possible reasons. Either they measure a different population on their different instruments (= population effect) or the instruments are calibrated differently (= real analytical bias).

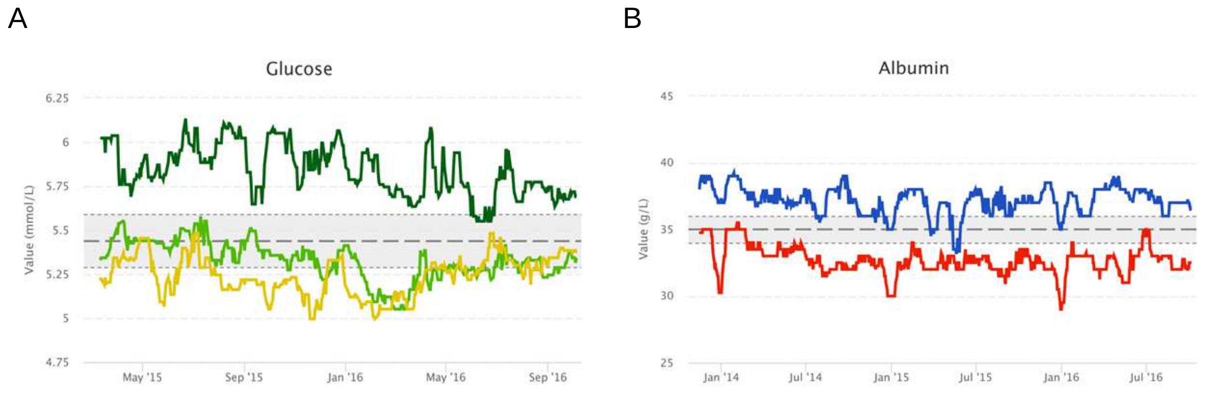


Figure 13. Examples of bias observations between different instruments in a single laboratory.

B) Bias assessment versus the peer group

By comparing the long-term median values of their own instruments (colored lines) with the peer median (black dotted line), laboratories can also use the Percentiler to check for possible bias issues between their own data and the peer group. These bias observations can have four possible reasons. Either a real analytical bias is present or the difference is caused by a mathematical factor, a population effect or by the use of a different method principle. In the latter case a laboratory might observe a bias because they use a less common method, e.g. creatinine Jaffé versus enzymatic. Not only the measurement method, but also the selected measurement matrix can cause a bias observation in the Percentiler (for example, plasma versus serum). Participants who send data on the measurement of plasma will usually observe a bias for potassium, inorganic phosphor, LDH and total-protein, because most participants send data on serum samples. In addition, some regions, e.g. The Netherlands and Scandinavian countries, have the habit of applying a factor to their measurement data, which will also result in a bias compared to the peer group. A number of examples, with bias-related issues not allocated to a real analytical bias, can be found in Figure 14.

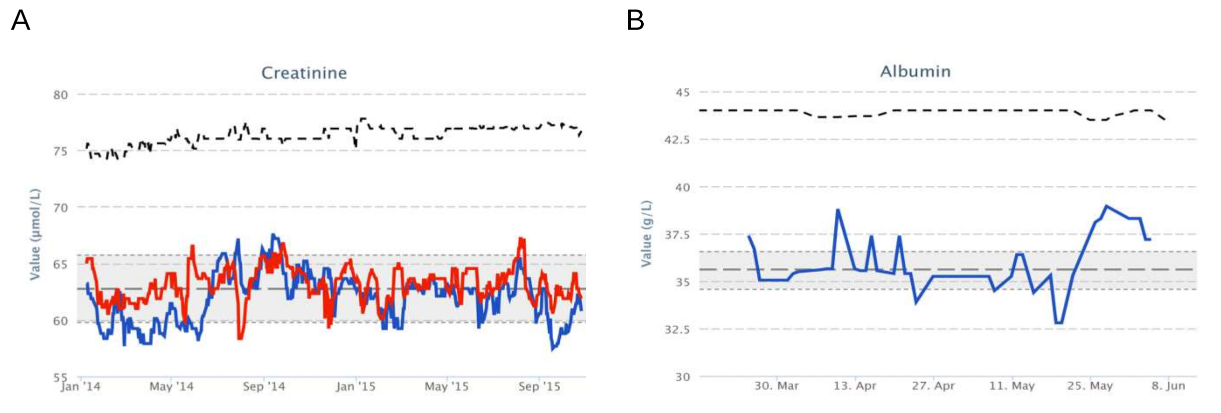


Figure 14. Examples of bias observations due to population effects and the use of different method principals. (A) Bias for creatinine to the Roche Cobas peer group due to a combination of a population effect and the use of a different method. Note, for creatinine two method principals exist, i.e. Jaffé or enzymatic, which give significantly different results. (B) Bias for albumine to the Roche Cobas peer group due to the use of a different method. This specific laboratory uses a immunoturbidimetric method to measure albumin, whilst most Percentiler participants use the bromcresol green (BCG) method.

To differentiate between a real analytical bias and a population effect, one cannot solely rely on Percentiler data. Laboratories can first cross-check the Percentiler observation with IQC data. In a second stage they can cross-check whether the same population is measured as for most Percentiler instruments (random sample assignment), or whether the laboratory preferentially measures, for example, inpatients. If one or several instruments would predominantly measure inpatients, one would expect, amongst others, a low albumin median, a low median for total-protein, and lower sodium values. In addition, the final proof can be made by measuring 20 left-over samples on two instruments, i.e., one instrument with and another without a bias versus the peer group. If the observed Percentiler bias is confirmed in this sample exchange experiment, one has proven the existence of an analytical bias. Two examples can be found in Figure 15. Note, for analytes with a narrow biological variation and a clear reference or peer target, such as electrolytes, it is easier to detect potential bias issues.

Mind, laboratories which use instruments belonging to a small peer group (< 20 instruments) can only compare their data with the median from all data, and not with the peer group. In that case an observed difference can also be caused by a lack of comparability between the different peer groups.

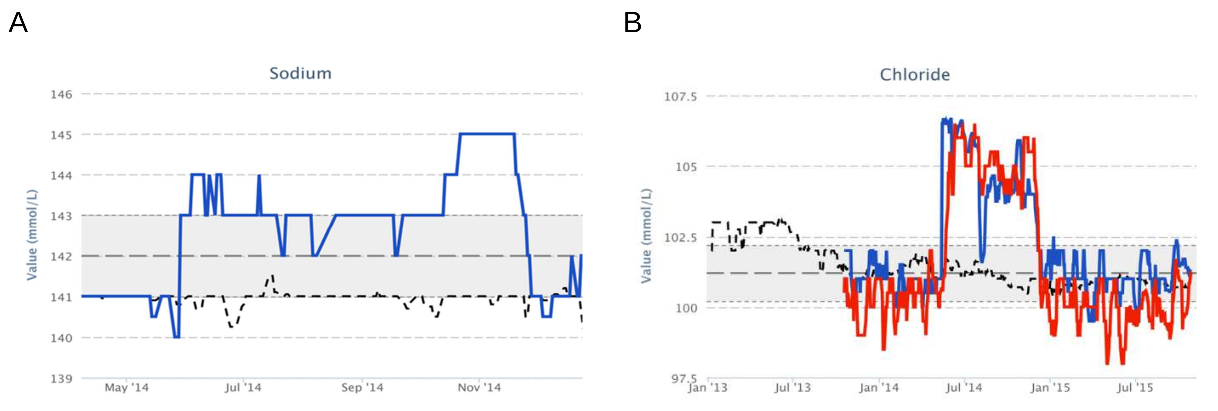


Figure 15. Examples of analytical bias observations. (A) Long-term bias compared to the peer for a sodium assay due to lot changes. The instrument measured at 143 mmol/L for a long time – which is a borderline significant bias compared to the long-term Vitros peer median of 141 mmol/L – and it even went up to 145 mmol/L (significant bias). (B) Long-term bias for a chloride assay compared to the Cobas peer. Due to electrode change this laboratory measured chloride medians around 105 mmol/L, whilst the Cobas peer is located at ~101 mmol/L.

C) Comparability assessment between peer groups

A final comparability assessment can only be performed by the project team (not by the individual laboratory). By extracting the median values grouped for one month from all Percentiler instruments we are able to assess the relative comparability status of the peer groups. This is done by presenting the monthly instrument medians in box- and whisker plots (peer specific). The box represents instruments within the 25th to 75th percentile; the whiskers extend to the minimum and maximum results. Therefore, these plots give a clear indication on the distribution of the instrument data per peer group, and it allows to compare those same peer groups. In addition, by creating box- and whisker plots for two time periods we are able to assess the robustness of the data. Figure 16 shows the data for GGT and chloride from March 2014 and July 2015. In general the observed patterns of the plots changed little over the course of the year for all analytes, at least for the peer groups which are sufficiently substantiated. In contrast, significant changes occurred for the small peer groups (n < 10), whilst other changes were restricted within the set bias specifications.

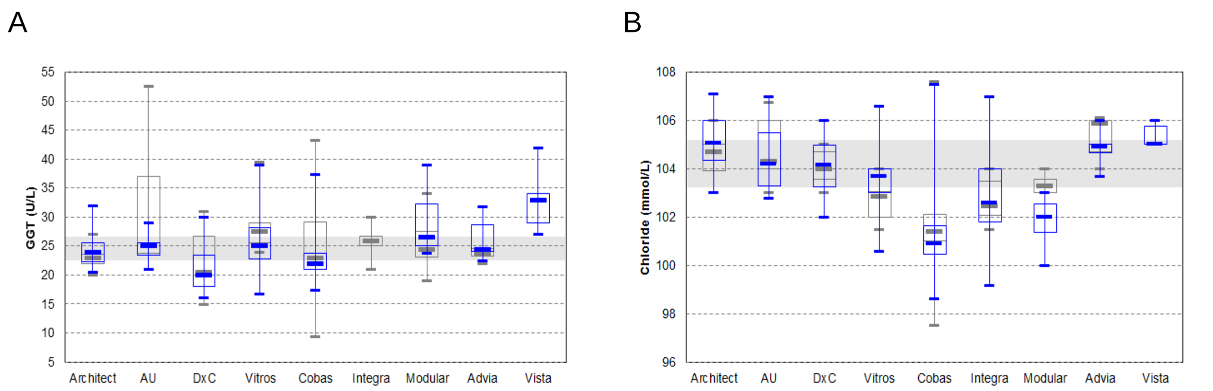


Figure 16. Box- and Whisker plots with the indication of peer group medians for (A) GGT and (B) chloride. The box represents instruments within the 25th to 75th percentile; the whiskers extend to the minimum and maximum results. The blue or grey bar in the box represents the monthly peer group median. Data from March 2014 are indicated in grey and set in the background; data from July 2015 are indicated in blue. To give an idea about the size of the dispersion, bias limits from the Percentiler are indicated with a grey shaded zone. These are constructed around the median value of the July 2015 data.

Normally, the comparability status of the different assays available on the market can only be reliably assessed with an EQA study making use of commutable samples. However, we tested whether comparability information from the Percentiler is reliable. To this end we compared the Percentiler peer group medians with the average of 20 samples from apparently healthy volunteers measured in our Master Comparison survey (see Chapter 4). Data in Figure 17 are presented as plots with indication of the Percentiler medians in blue, and Master Comparison 2014 sample averages in red (for each peer group). Note, the sample averages for the Master Comparison study were normalized to the patient medians. This was necessary because stratification for outpatients may not completely exclude the influence of diseased persons (mostly those having chronic diseases such as diabetes or kidney disease can significantly influence certain analytes, such as creatinine, glucose, uric acid, etc.).

As obvious from Figure 17, the relative comparability between the peer groups was very similar in the Percentiler and Master Comparison data, for most analytes and peer groups. Observed differences could be explained by the occurrence of small peer groups in either the Percentiler or the Master Comparison survey. This supports the potential of the Percentiler to study relative assay biases at median concentrations. However, the need for adaptation of the Master Comparison data to the Percentiler data may indicate that assessment of "absolute" bias is more challenging, in particular, for the enzymes. The latter (assessment of "absolute" bias) may require a more stringent outpatient stratification or a selection of laboratories which mainly serve the general population.

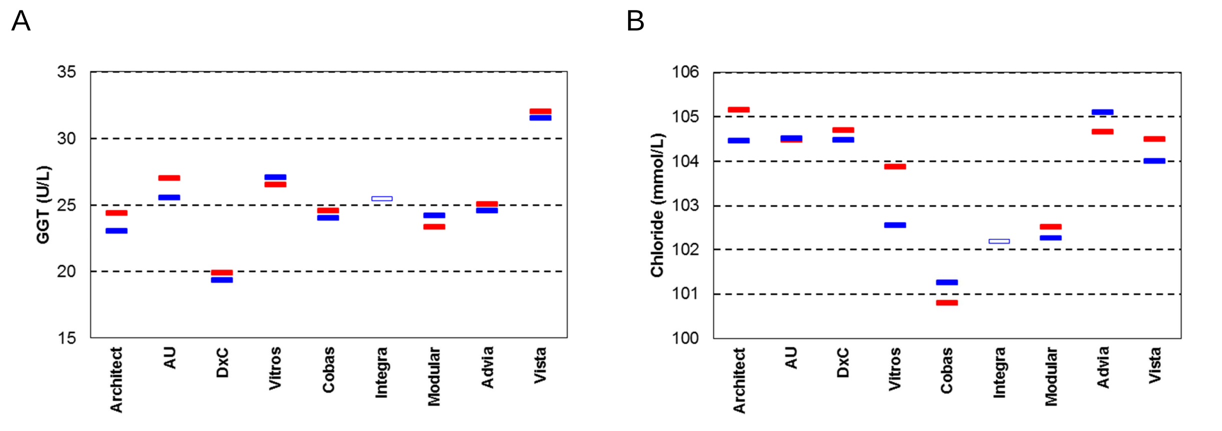


Figure 17. Plots with indication of the Percentiler medians in blue (from May 2014), and Master Comparison 2014 sample averages in red (for each manufacturer/peer group); for (A) GGT and (B) chloride. The sample averages for the Master Comparison study were normalized to the patient medians.

Link with the Committee for Standardization of Thyroid Function Tests

In a later stage, FT4 and TSH were added to the Percentiler and Flagger database. This was done in the framework of the Committee for Standardization of Thyroid Function Tests (C-STFT), which aimed at using the Percentiler for monitoring the stability of the assays which participate in their standardization efforts. Indeed, clinicians diagnose and follow thyroid dysfunction based on TSH and FT4 testing. However, the current lack of comparability between assays limits the optimal use of these laboratory data. The IFCC gave a mandate to the C-STFT to resolve this limitation by standardization of FT4 and harmonization of TSH. However, before implementation of the technical recalibration, they were furthered by the Food and Drugs Administration (FDA) to develop a tool to assess the sustainability of the new calibration basis. C-STFT began to use the Percentiler and the Flagger, with the intention to assess their utility for this purpose. Current data suggest the suitability of both applications to document the sustainability of the calibration basis in the post-standardization phase. For more information we refer to a publication from the doctoral work from Linde De Grande (20).

7. Conclusion on the Percentiler

The Percentiler has added value compared to the limitations of traditional QC. First, there are little additional costs involved (no measurement of additional samples). Second, the materials used for the Percentiler are “as commutable as they possibly can be”, i.e. patient samples. And, third, the Percentiler is an online, continuous quality monitoring tool, in contrast to the low frequency surveys typical for EQA. It provides the mid- to long-term analytical stability assessment against realistic specifications guided by biological variation; if this was not feasible, state-of-the-art performance limits were adopted. Observations from the Percentiler graphs show that test variability due to confirmed calibration and lot changes is still a significant problem for both laboratories and manufacturers.

Naturally, the Percentiler also has limitations. First, patient data have of course a much higher variation then single-concentration IQC samples. Therefore, the Percentiler’s utility is the better the higher test volumes are achieved and the more carefully stratification for outpatient results is done. The latter might be difficult to achieve because different hospital have different lab policies concerning (out)patient stratification and the LIS market is highly fragmented. Second, patient percentile monitoring is typically restricted to the median of the results because monitoring at “outer percentiles” (e.g., at 2.5%- or 97.5%-percentile) results in much higher variation.

Concluding, patient percentile monitoring for mid-to-long-term quality management is particularly attractive for laboratories as well as manufacturers, because it has the potential to build a continuous, global evidence base on IVD test stability (and relative comparability). It gives evidence about stability of performance and the reasons for assay variation (manufacturer, lot-to-lot, calibration, instrument). In addition, it helps with the establishment of realistic quality specifications derived from state-of-the-art performance, and it provides a basis for comparison across manufacturers.

References

1. Kazmierczak SC. Laboratory quality control: using patient data to assess analytical performance. Clin Chem Lab Med 2003;41(5):617-27.
2. Houwen B, Duffin D. Delta checks for random error detection in hematology tests. Lab Med 1989;20:410-7.
3. Hoffmann RG, Waid ME. The "average of normals method" of quality control. Am J Clin Pathol 1965;43:134-41.
4. Kilgariff M, Owen JA. An assessment of the "average of normals" quality control method. Clin Chim Acta 1968;19:175-9
5. Douville P, Cembrowski GS, Strauss JF. Evaluation of the average of patients: application to endocrine assays. Clin Chim Acta 1987;167:173-85.
6. Lott JA, Smith DA, Mitchell LC, Moeschberger ML. Use of medians and "average of normals" of patients' data for assessment of long-term analytical stability. Clin Chem 1996;42:888-92.
7. Westgard JO, Frederick IA, Smith FA, Mountain PJ, Boss S. Design and assessment of average of normal (AON) patient data algorithms to maximize run lengths for automatic process control. Clin Chem 1996;42:1683-8.
8. Stepman HC, Stöckl D, Stove V, Fiers T, Couck P, Gorus F, Thienpont LM. Long-term stability of clinical laboratory data: sodium as benchmark. Clin Chem 2011;57:1616-7.
9. Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P, et al. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.
10. Stepman HC. Analytical quality and its effect on medical decisions of laboratory tests for the management of chronic kidney disease – calcium, albumin, phosphorus, 25-hydroxyvitamin D and parathyroid hormone. Doctoral Dissertation 2013.
11. Westgard QC. Desirable specifications for total error, imprecision, and bias, derived from biologic variation. https://www.westgard. com/biodatabase1.htm (accessed January 2014)
12. Nordic Reference Interval Project (NORIP). http://pweb.furst.no/norip/ (accessed January 2014).
13. Robert COH, Thomas RH. Causes and evaluation of mildly elevated liver transaminase levels. American Family Physician 2011;48:1003-8.
14. Aragon G, Younossi ZM. When and how to evaluate mildly elevated liver enzymes in apparently healthy patients. Cleveland Clinic Journal of Medicine 2010; 77:195-204.

15. Asirvatham JR, Moses V, Bjornson L. Errors in potassium measurement: a laboratory perspective for the clinician. N Am J Med Sci 2013;5(4):255-9.

16. Lippi G, Salvagno GL, Montagnana M, Brocco G, Guidi GC. Influence of hemolysis on routine clinical chemistry testing. Clin Chem Lab Med 2006;44(3):311-6.

17. Maes M, Mommen K, Hendrickx D, Peeters D, D'Hondt P, Ranjan R et al. Components of biological variation, including seasonality, in blood concentrations of TSH, TT3, FT4, PRL, cortisol and testosterone in healthy volunteers. Clin Endocrinol (Oxf) 1997;46:587-98.

18. Leppaluoto J, Sikkila K, Hassi J. Seasonal variation of serum TSH and thyroid hormones in males living in subarctic environmental conditions. Int J Circumpolar Health 1998;57(Suppl 1):383–5.

19. De Grande LA, Goossens K, Van Uytfanghe K, Halsall I, Yoshimura Noh J, Hens K, Thienpont LM. Using "big data" to describe the effect of seasonal variation in thyroid-stimulating hormone. Clin Chem Lab Med 2016; 55(2):e34-e36.

20. De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru M, Thienpont LM, for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta 2016. doi: 10.1016/j.cca. 2016.04.032. [Epub ahead of print].

Article 1. The Percentiler

Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make "Big Data" work

Kenneth Goossens[1], Katleen Van Uytfanghe[1], Patrick J Twomey[2], Linda M Thienpont[1*], and Participating laboratories

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

[2]Department of Clinical Chemistry, St Vincent's University Hospital, Dublin 4, Ireland

Introduction

Laboratory data has the potential for substantially aiding the development, implementation, and management of public healthcare policies. It can create public awareness of the importance of maintaining a healthy lifestyle as well as reacting early to signals of health problems. As such, it can indirectly contribute to reduce the burden of healthcare expenses. The drive to focus on improved exploitation of laboratory data typically comes from financial pressures, such as the steady increase in health-care expenses in the US during the last 20 years. Such expenses now represent 17.6% of the gross domestic product and nearly $600 billion more than the expected benchmark for a nation of the size and wealth of the US (1). An additional impetus to transform the laboratory landscape comes from the information technology (IT) revolution, offering, among others, the opportunity to create reliable and accessible "Big Data" (2, 3). Nevertheless, the "big bang" for the active role of IT in healthcare policy came in the US from legislation "The Health Information Technology for Economic and Clinical Health (HITECH) Act" (4) and the push by the government to adopt electronic health records (EHRs) (5-8). The research firm Frost & Sullivan predicts that use of advanced health data analytics solutions in hospitals will increase to 50% adoption by 2016 (9). This may create enormous business opportunities, for example, the Washington Post reported the inclusion of "as much as $36.5 billion in spending to create a nationwide network of electronic health records" (10). However, big spending should be justified by big savings. Indeed, according to a report from McKinsey & Company, the largest managed care organization in the US (Kaiser Permanente), reported that their "Big Data" strategy has saved the organization $1 billion in reduced office visits and lab testing (1).

If data is the new gold, then access to data is going to be key to insights (2), but key is also the expertise of the laboratory to ensure the reliability of the data as well as its safe and efficient use. While promises are sky-high, EHR is not without risks, especially in the start-up phase. In this regard "The November 2011 Institute of Medicine report, Health IT and Patient Safety: Building Safer Systems for Better Care" noted that the lack of empirical data on the nature and prevalence of EHR system-related adverse patient events makes it challenging to determine the extent of the associated risks to patient safety (11). If one takes this note seriously, the question is how to minimize risks? Can a small non-US based group play a role in this highly innovative, competitive, multi-billion dollar environment? We believe yes, because we are convinced that efforts to take care of the quality of the input data will improve the quality of the output. The utility of the EHR is often overlooked as a result of laboratory diagnostics sometimes giving significantly different results for the same patient sample, even for the simple, high-volume clinical chemistry tests (12-14).

In an attempt to illustrate this limitation and more importantly to do something about it, we describe "The Percentiler" project, which is part of our overarching "Empower" project introduced elsewhere (15). In essence, it provides quasi real-time access to patient medians across laboratories and manufacturers. This data can serve as a "clearinghouse" for potential future EHR applications, such as the retrieval of laboratory data for epidemiological or toxicological research on national or global scale, long-term follow-up of chronic diseases, or linking laboratory data to mortality risk (16, 17).

Materials and methods

Participants and participation process

Participating laboratories are globally distributed. They range from medium-sized to big hospital laboratories, but also include private laboratories (for the current list of participants, see www.stt-consulting.com, Empower tab). When a laboratory declares its intention to join, we provide it with the information about the IT requirements for sending data, together with a request for a test e-mail. One of our project team controls the test-mail, maps the data and verifies error-free transmission into our database. If successful, we continue this verification for a while before giving the definitive Percentiler e-mail address. Subsequently, data transfer either occurs automatically and on a daily basis (depending on the Laboratory Information System (LIS)) or is done in manual batches. After sending data for 6 to 8 weeks, the participating laboratory receives its login information, which gives access to the graphical presentation of its data via a user interface. Data is assessed by peer group: typically 10 or more laboratories using the same test system. Participation is free of charge. Furthermore, all LIS solutions for automatic median calculation and data transfer are provided at no or minimal cost and without running costs.

Data

We collect instrument-specific daily medians calculated from outpatient results of 20 commonly measured analytes in serum or plasma: albumin, alanine aminotransferase (ALT), alkaline phosphatase (ALP), aspartate aminotransferase (AST), calcium, chloride, C-reactive protein (CRP), creatinine, γ-glutamyl transferase (GGT), glucose, inorganic phosphorus (phosphate), lactate dehydrogenase (LDH), magnesium, potassium, sodium, total-bilirubin, total-cholesterol, total-protein, urea, and uric acid (urate).

Data coding and transfer to a database

Data coding comprises 7 attributes each separated by "semicolon": laboratory identification (Lab ID); date (e.g., 02/01/2014); instrument identification (Instr ID); code for outpatients (e.g., OUT); test name (e.g., CA for calcium); test unit (e.g., mmol/L); median (e.g., 2.35). The laboratories can retrieve these attributes directly from the LIS and adopt the used mnemonics. The only requirement is for the laboratories to organize the data in a table according to the format below:

Lab ID; 02/01/2014; Instr ID; OUT; CA;mmol/L; 2.35

Lab ID; 02/01/2014; Instr ID; OUT; NA; mmol/L; 141

Lab ID; 02/01/2014; Instr ID; OUT; CL; mmol/L; 102.5

Data must be sent electronically to our project-specific e-mail address either as i) an e-mail embedded table, ii) an e-mail attached EXCEL-file; or iii) an e-mail attached text-file. Retrieval of data and electronic exporting is done either automatically (by features available in a specific LIS), or manually. Automatic solutions send the data daily, while manual solutions operate in a batch fashion with the data manually extracted weekly or monthly and manually sent by e-mail.

Database

The software for data downloading from the e-mail, transfer into a MySQL database, and the development of "The Percentiler" application and user interface was programmed by Bruno Neckebroek (Zwijnaarde, Belgium). Data from the individual laboratories are "mapped" by the STT/UGent project team to common analyte names, units, and instrument names and other technical details.

Data analysis/User interface

The database is fully accessible to the STT/UGent project team, who investigate laboratory and peer group data for bias and trends. Critical observations are communicated in the first instance to the laboratories concerned. They are also shared with instrument vendors, and regularly, with the whole group of participants. It is important to note that the identity of the laboratory is not disclosed to a third party under any circumstances. The user interface (accessed via a specific login at https://thepercentiler.be) only gives the laboratories access to their own data (login with username and password). Investigation of data is possible on-line. However, if detailed off-line analysis is preferred, the data can be downloaded into Excel. Other functionality in the user interface allows a downloadable chart of the moving median in time (laboratory and peer), and a table with summary statistics (bias, robust CV) for each analyte. The selection possibilities include i) n for calculation of the moving median (n = 5, 8, 16); ii) time window; iii) inclusion/exclusion of weekends. When participants report medians for two or more instruments, an instrument-specific color code is used in the charts. The assessment of the stability of laboratory testing is done against desirable bias limits from biological variation, at least for the analytes for which state-of-the-art performance allows this. However, the maximum bias limit is set to ~10%. The limits are visualized in the charts by a gray zone, and violations >1 week are considered significant. For more detailed information, the reader is referred to the demo version of "The Percentiler" (https://thepercentiler.be, login: demolab, password: demo1234).

Partners

The Royal Belgian Society of Clinical Chemistry scientifically supports the project. The assistance from several LIS vendors in providing solutions for automatic data calculation, retrieval, and electronic sending greatly contributed to the practical realization of the project. Further support is received from the Belgian representatives of the main in vitro diagnostic manufacturers (see www.stt-consulting.com, Empower tab, for LIS and manufacturer information).

Results

Participation and reporting

Currently (December 2014), 124 laboratories participate with ~250 instruments, distributed over the following peer groups: Advia (n = 8); Architect (19); AU (13); Cobas (153); Integra (3); Modular (11); Synchron (11); Vista (6); Vitros (26). Participation is global (see Fig. 1), however, most of the current participants come from Belgium.

As of December 2014, several reports have been produced that address i) the general features of the project, ii) peer group observations; iii) synergisms between "The Percentiler" and our dedicated EQA surveys, the so-called "Master Comparisons (MCs)". These reports are accessible at www.stt-consulting.com (under the Empower-Tab). Exemplary observations are presented below.

Test comparability

Table 1 presents the peer group data for the 5 most prominent instrument types (Architect, AU, Cobas, Synchron, and Vitros). It also shows the %-differences between the 2 peer groups that agree the least (column A), gives the 2-times desirable bias limits from biological variation (2 times because of comparison with the peer group differences) (column B), and the ratio of column A/column B. Several analytes exhibit ratios ≥2, indicating a significant lack of comparability between manufacturers. This holds particularly true for enzymes due to the use of different method principles (transaminases with and without pyridoxal phosphate activation; LDH forward and reverse reaction). Significant differences are also observed for albumin, total-bilirubin, chloride, CRP, glucose, and magnesium.

The observations correspond closely to those of the MCs surveys (19) (for example, chloride: Fig 2). Data is presented as plots with the patient medians in blue, and sample means (in the 2014 MCs survey) in red (for each device). Note that the sample means for this survey were normalized to the patient medians. Except for Vitros, the relative differences between the instruments are quite similar in the MC and the Percentiler application, with Cobas showing the lowest values.

Test stability

Test stability is assessed against desirable bias limits derived from biological variation or from state-of-the-art performance (Table 2). The latter limits were applied for analytes with small biological variation where state-of-the-art performance cannot meet the challenging bias limits from biological variation (albumin, calcium, glucose, potassium, magnesium, and sodium). However, the maximum bias is set to ~10%. Limits were set in units rather than percentage, e.g., for CRP 0.2 mg/L which results in a limit of 11%. Typically, >90% of the laboratories showed a stable performance over 1 year within the chosen limits. Peer group influences could be investigated for only 5 instruments (Architect, AU, Cobas, Synchron, and Vitros). Among them, Vitros scored lowest with only ~80% of the laboratories achieving stable performance over the year. Six examples can be found in Figure 3 to 5.

Figure 3 shows the time course of the chloride median in a laboratory for ~2 years. Stability and comparability of the 2 instruments were excellent (long-term median ~103 mmol/L, stability ±1 mmol/L), with the exception of a small shift in July

Table 1. Peer group medians, %-differences of the 2 most deviating peer groups (column A), 2-times the desirable bias from biological variation (column B), and ratio of column A/column B for the 5 major peer groups present in "The Percentiler" application; ratios A/B ≥2 indicated in bold. (Note, data on biological variation shows values as they were at December 2014.)

Analyte	Unit	Architect	AU	Synchron	Vitros	Cobas	Ratio: A/B	A: Δ (%)	B: 2 * Bias
ALB	**g/L**	41.0	42.6	40.0	40.5	44.0	**3.7**	9.6	2.6
ALP	**U/L**	69.5	83.0	70.4	75.8	74.0	1.4	18	12.8
ALT	**U/L**	18.7	20.0	20.0	27.4	18.0	**2.0**	45	22.8
AST	**U/L**	20.0	22.6	21.0	26.0	20.0	**2.5**	27	10.8
BIL-T	**µmol/L**	9.5	10.0	11.5	9.2	6.8	**2.2**	50	22.8
CA	**mmol/L**	2.37	2.38	2.36	2.39	2.37	0.8	1.3	1.6
CHOL	**mmol/L**	5.06	4.89	4.95	4.67	4.78	1.0	8.0	8.0
CL	**mmol/L**	105.1	104.1	104.7	102.5	100.5	**4.4**	4.4	1.0
CREA	**µmol/L**	71.6	74.9	73.5	75.1	76.0	0.7	5.9	8.0
CRP	**mg/L**	2.1	2.5	3.1	6.2	2.7	**2.8**	124	43.6
GGT	**U/L**	25.0	25.0	21.5	27.0	22.0	1.1	23	21.6
GLUC	**mmol/L**	5.35	5.62	5.83	5.50	5.27	**2.3**	10	4.4
K	**mmol/L**	4.35	4.30	4.15	4.30	4.38	1.5	5.4	3.6
LDH	**U/L**	191.5	183.0	178.5	489.5	172.5	**15.2**	131	8.6
MG	**mmol/L**	0.84	0.83	0.86	0.79	0.82	**2.4**	8.5	3.6
NA	**mmol/L**	140.0	139.0	139.0	140.5	140.6	1.8	1.1	0.6
P	**mmol/L**	1.09	1.07	1.17	1.21	1.13	1.9	12	6.4
PROT	**g/L**	71.0	70.1	69.3	72.2	69.5	1.7	4.1	2.4
UREA	**mmol/L**	5.32	5.59	5.05	5.66	5.00	1.1	12	11
UA	**µmol/L**	321	312	320	315	324	0.4	3.8	9.8

Albumin (ALB), alanine aminotransferase (ALT), alkaline phosphatase (ALP), aspartate aminotransferase (AST), calcium (CA), chloride (CL), C-reactive protein (CRP), creatinine (CREA), γ-glutamyl transferase (GGT), glucose (GLUC), inorganic phosphor (P), lactate dehydrogenase (LDH), magnesium (MG), potassium (K), sodium (NA), total-bilirubin (BIL-T), total-cholesterol (CHOL), total-protein (PROT), urea, and uric acid (UA). *Calculation example for ALP: Δ = 100*[83 (AU) –69.5 (Architect)]/75.8(Average of all) = 18%.

Table 2. Bias limits set for "The Percentiler" project and compared to desirable bias from biological variation. In the application, the limits are set on a unit basis and are considered violated if the bias lasts >1 week; state-of-the-art limits are in bold.

	Biology	Percentiler	Biology	Percentiler	
	(%)	**(%)**	**(unit)**	**(unit)**	**Unit**
ALB	1.43	**2.4**	0.62	1	g/L
ALP	6.72	6.8	5.0	5	U/L
ALT	11.48	9.5	2.2	2	U/L
AST	6.54	6.5	1.4	1.5	U/L
BIL-T	8.95	10.0	0.73	1	µmol/L
CA	0.82	**1.7**	0.019	0.04	mmol/L
CHOL	4.1	3.8	0.20	0.2	mmol/L
CL	0.5	**1.0**	0.51	1	mmol/L
CREA	3.96	3.9	3.0	3	µmol/L
CRP	21.8	**9.6**	0.40	0.25	mg/L
GGT	11.06	9.1	2.4	2	U/L
GLUC	2.34	**3.1**	0.12	0.15	mmol/L
K	1.81	**2.4**	0.08	0.1	mmol/L
LDH	4.3	4.6	7.7	8	U/L
MG	1.8	**3.0**	0.015	0.025	mmol/L
NA	0.23	**0.7**	0.32	1	mmol/L
P	3.38	4.4	0.038	0.05	mmol/L
PROT	1.36	1.4	0.95	1	g/L
UREA	5.57	6.0	0.30	0.3	mmol/L
UA	4.87	4.8	15.6	15	µmol/L

Abbreviations used for the different analytes are identical as in Table 1. Note, values were updated to the most recent numbers (October 2016).

2013. Similar good stability was observed for sodium and calcium (see Figures 5A and 5B).

Figure 4 shows the time course of the ALT median in a laboratory over more than 1 year. The data are characterized by drifts and calibration/lot shifts, with maximum values at ~37 U/L and minimum values at ~22 U/L. Other examples of test instabilities for chloride and creatinine can be found in Figures 5C and 5D. Problems for the chloride test have been observed several times, mainly due to shorter electrode lifetimes than those claimed by manufacturers.

Laboratory bias

To date, assessment of laboratory bias is done with caution as only the Cobas peer group target is calculated from sufficient instruments. In addition, more experience needs to be gained about the potential influence of population effects. Nevertheless, some grossly deviating results have been reported by individual laboratories, findings that were confirmed by sample exchange experiments.

Pre-analytical effects

Pre-analytical effects are seen in certain private laboratories. Follow up determined that this is due to a considerable delay between sample collection and processing. For potassium, this led to high values in winter and normal values in summer. The opposite pattern was observed for LDH medians.

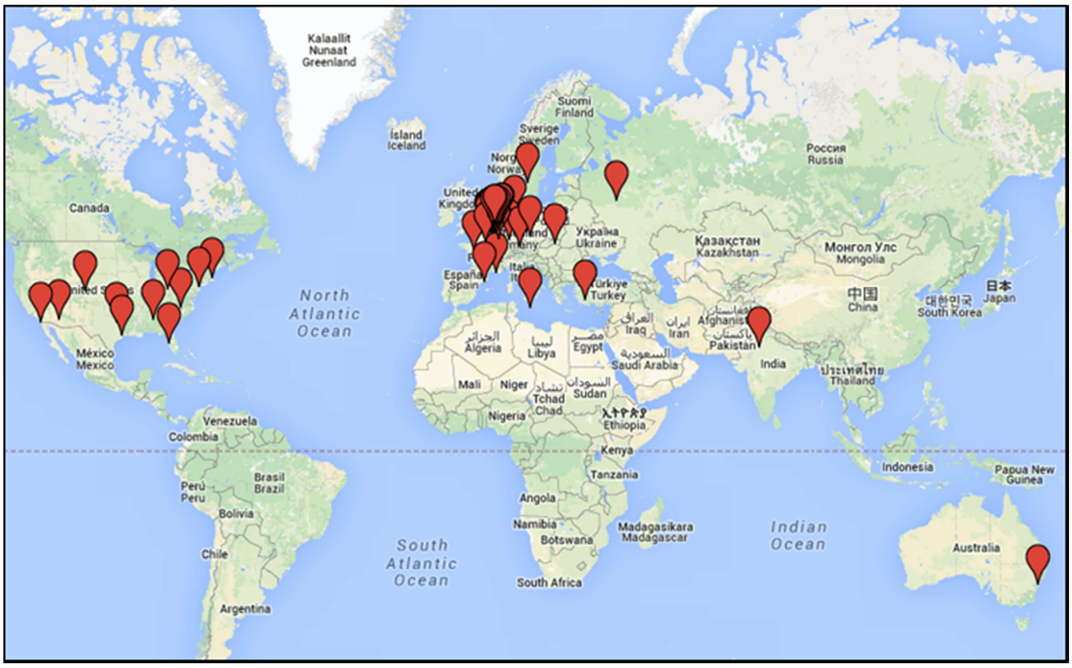


Figure 1. Global geographic distribution of the participants in the "The Percentiler" project (December 2014)

Discussion

We showed that "The Percentiler" project provides a real-time monitoring tool for test comparability and stability across laboratories and manufacturers. Typical attributes make the approach different compared to conventional external quality assessment/proficiency testing [20]. It uses samples that are "as commutable as can be" and data that are already available in the laboratory at little extra costs. It provides the mid- to long-term analytical

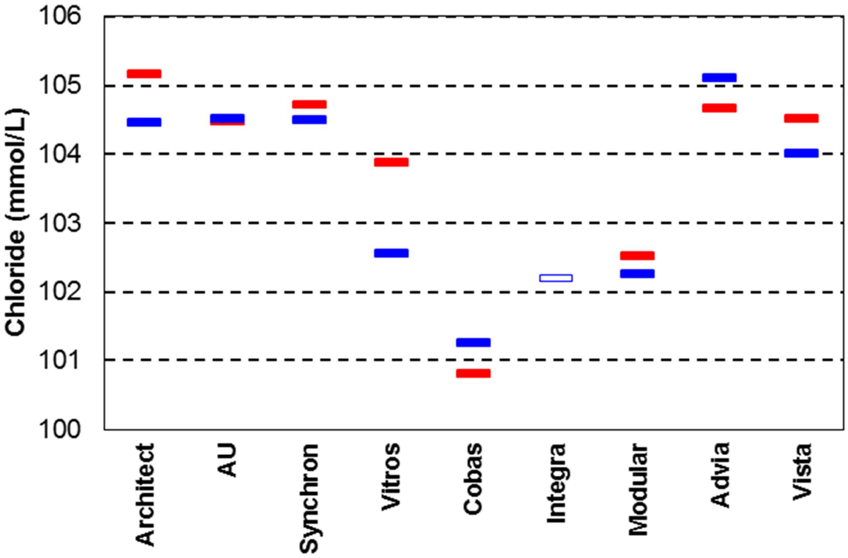


Figure 2. Peer group comparability for chloride in “The Percentiler” and “Master Comparisons” 2014 survey (patient medians in blue, MCs sample means in red)

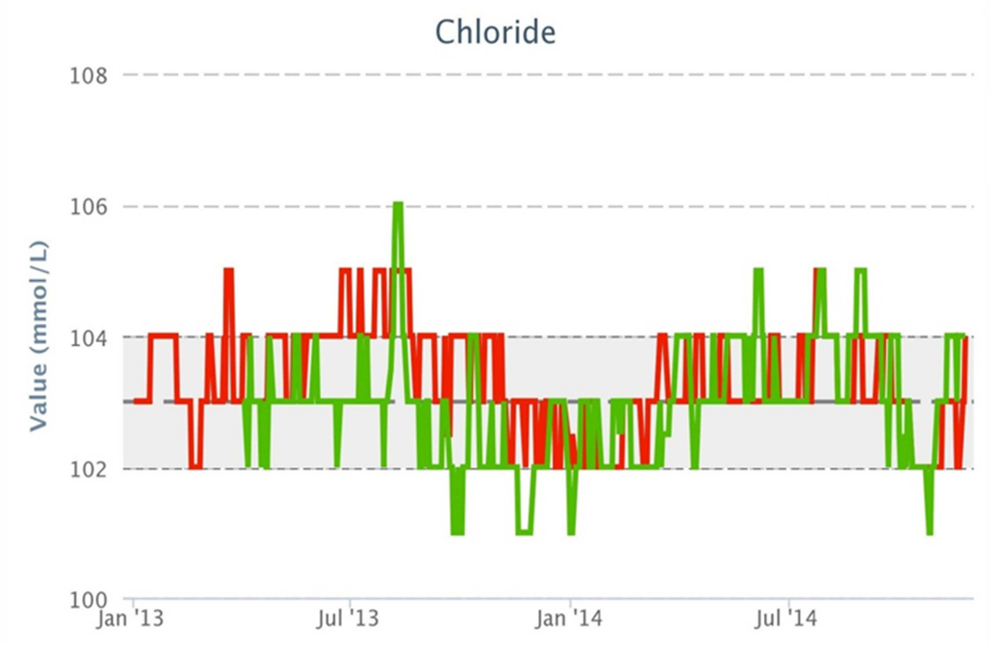


Figure 3. Time course of the chloride moving median in a laboratory with low analytical variability and excellent instrument comparability over ~2 years.

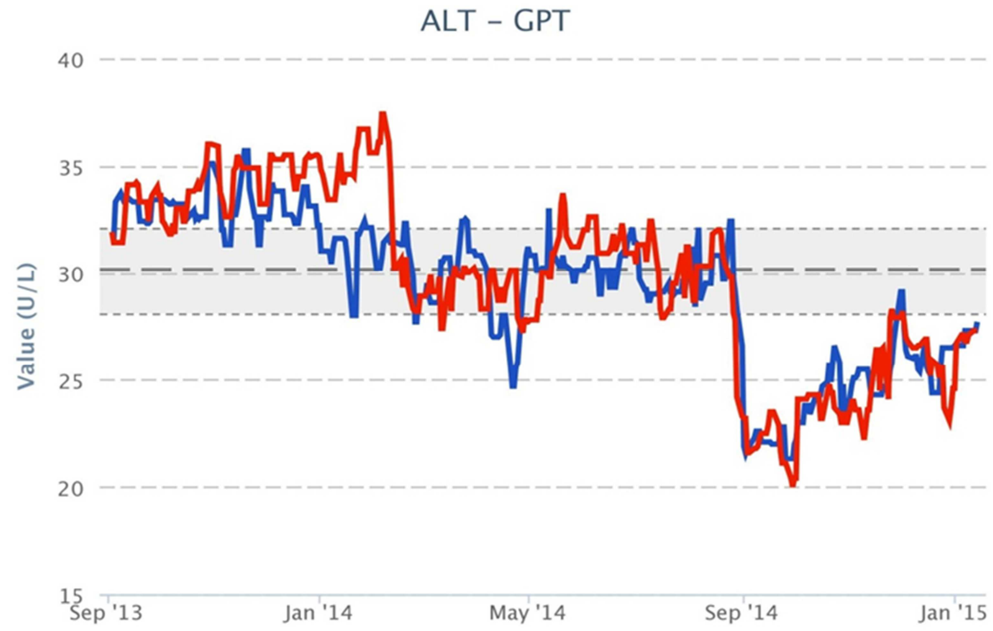


Figure 4. Time course of the ALT moving median in a laboratory with high analytical variability due to calibration and lot effects.

stability assessment against realistic limits guided by biological variation [18]; if this was not feasible, state-of-the-art performance limits were adopted. The application is useful for all type of laboratories, from small-sized (for example, 250 beds) to high-throughput ones. The only difference is that the former have to choose a higher n (8 to 16) for calculation of the moving median, and hence, will see instability/changes later than the high-throughput laboratories. Global interest is demonstrated from the geographic distribution of the participants, which is another advantage over many nationally operating external quality assessment/proficiency schemes.

Monitoring of outpatient medians and instrument-based peer groupings demonstrates significant differences among manufacturers for many of the 20 tests we currently monitor (for details, see Table 1, ratio A/B). The observations are concordant with the differences we have seen in our previous MCs surveys, in particular, for commonly measured enzymes, albumin, phosphate, total-bilirubin, and chloride [12, 13, 19]. We admit that the observed differences may be influenced by instrument-based peer groupings without accounting for the method principle employed, e.g., for creatinine, on certain platforms the Jaffe assay is installed, while on others the enzymatic assay is utilized. From this point of view, it might be beneficial to consider this distinction. On the other hand, we think it also makes sense to point to the fact that different assays for measurement of common analytes still lead to non-

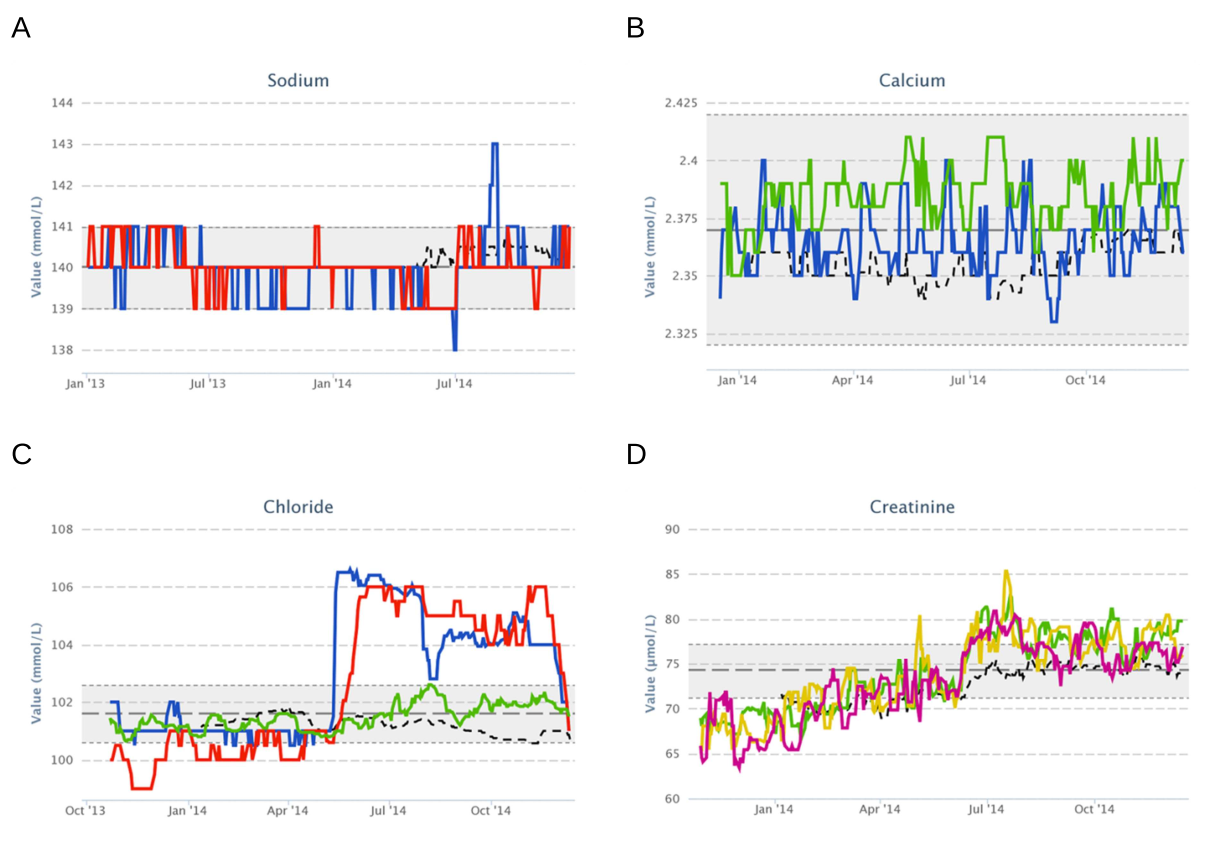


Figure 5. Selected time course examples from “The Percentiler” application (stability limits in gray; peer group indicated as broken line). (A) Sodium moving median in a laboratory with a 2-year lasting “1 mmol/L stability” and excellent instrument comparability. (B) Calcium moving median in a laboratory with >1 year excellent stability and instrument comparability. (C) Chloride moving median in a laboratory with robustness problems of the ion selective electrode. (D) Creatinine moving median in a laboratory with long-term drift and shifts, partly due to lot changes.

comparable results and thus need better standardization. Nevertheless, we will consider peer grouping based on system and method principle, however, interpretation at that level will require many more participants, as well as modifications in the LIS data transfer logic. Other observations show that test variability due to confirmed calibration and lot changes is still a significant problem for several laboratories and manufacturers, which holds particularly true for certain enzymes (ALT, AST). The project further confirms that the ion selective electrode technique still has robustness issues for chloride measurement (occasionally also for sodium), which, typically, can be solved by replacement of the electrodes earlier than claimed by the manufacturers [21]. It is comforting that many laboratories are able to achieve a long-term stability (>1 year) within the desirable bias limits inferred from biological variation data for most of the tests, as also shown in other studies [22-24]. For analytes with narrow biological variation, they accomplish a stable performance of 1 mmol/L for sodium and chloride, 1 g/L for total-protein and albumin, and 0.05 mmol/L for calcium. In certain private laboratories, significant pre-analytical problems due to delayed sample processing jeopardize the reliability of potassium and LDH testing. Laboratory bias is generally more difficult to uncover from patient medians because of the potential influence of a lower sample size and higher population variation in certain laboratories. Indeed, in low- to medium-sized laboratories the effects from biological variability (e.g., gender, race and age), or inclusion on certain days of data for special patient categories (e.g., dialysis or oncology patients often are registered as outpatient in the LIS) might not be leveled off as they do in high throughput laboratories. Nevertheless, we were able to confirm certain cases where we suspected laboratory biases. This was possible because the laboratories concerned were eager enough to perform sample exchanges with other participants. Therefore, at the current stage of this project, we recommend to our participants, especially those from small- and medium-sized hospitals, to investigate laboratory bias by participating in peer group internal quality control programs, while using the monitoring of patient medians for longer-term quality management. Notwithstanding this, it is worth noting that the project may benefit from better information concerning the population variation. Once it has grown sufficiently, we will try to expand our information base, however, again, this will require adaptations for the data transfer logic by the LIS companies.

In this project, we use to share critical observations concerning test instability over time by taking contact with individual laboratories and inviting them to do further investigations. As a result, we could confirm that biases/trends can be frequently attributed to lot changes or calibration events, or, as discussed above, that the laboratory itself was the cause. Likewise we share peer group observations with the respective laboratories and the relevant manufacturers. If the sole achievement of this project is just to evoke a response from the concerned parties, “The Percentiler” project will be of significant benefit to all stakeholders involved in laboratory medicine (laboratories, manufacturers, physicians, society, patients).

The project provides the individual laboratory with a direct, real-time quality indicator for its analyses of patient samples, not confounded by commutability issues [25]. There is no minimum number of data points required, however, the laboratory should be aware that the lower the number, the higher the resulting variability of the medians. As stated before, this requires that small-sized laboratories choose a higher n for calculation of the moving median. Despite this, they still will be able to investigate the stability of their analytical performance, but they will see changes later than the high-throughput laboratories. Together with data from internal quality control, the project helps the laboratory to build evidence about mid- to long-term stability of assays/performance, as well as of the reasons for variation: laboratory performance or manufacturer performance (for example, lot-to-lot-variation) [22, 23]. Thus, it generally strengthens the overall quality management/quality assurance system. This evidence is backed-up by information from other laboratories using the same assay. Knowledge of the above effects can also strengthen the laboratory/manufacturer dialogue. Cross comparison between manufacturers allows insight into the mid- to long-term performance of other tests and may aid in selection of future test systems. Monitoring of patient data also creates a tool for developing realistic quality goals [22, 24].

The project provides manufacturers with real-time data about field performance of their instruments and application issues in particular laboratories. Because the employed data are from analysis of patient samples, discussion about performance of third party quality assurance samples becomes obsolete. Also, the possibility offered by the project to see peer group trends and/or to be informed about peer group problems, allows manufacturers to verify issues more easily than by communication with single laboratories. Cross comparisons with other peer groups allow the identification of state-of-the-art performance and give impetus for test improvement when significant sub-standard performance is observed for certain tests.

The project may improve the physician/laboratory interface because communication is about patient results and not, for example, about trends observed for quality control samples. At the same time, it helps physicians to develop realistic expectations about inevitable variability of laboratory data. Ultimately, interoperable and stable laboratory data facilitate and improve care that physicians can offer to their patients.

The project is beneficial for the society, and finally for the patient. The data can serve as a “clearinghouse” for potential future EHR applications. This is particular important when laboratory data are intended for use in epidemiological or toxicological research on a national or global scale. Continuous evidence for test comparability and stability is of paramount importance for long-term follow-up of chronic diseases (diabetes, thyroid-, kidney- and cardiovascular disease) and linking laboratory data to mortality risk [e.g., 16, 17]. However, only when laboratory data is interoperable, can it fulfill the promise of “Big Data”, such as cost reduction, improvement of current practices, and creation of new insights for early diagnosis and long-term public health care management.

Of course, the monitoring of patient data also has limitations: the greater the test volume and the more careful the selection of outpatient results, the greater the utility of the data. In addition, monitoring is most effective at the median level because “outer percentiles” (for example, 2.5th or 97.5th percentiles) exhibit a significantly higher population variation than do central tendencies such as the median.

Conclusion

“The Percentiler” project has the potential to build a continuous, global evidence base on in vitro diagnostic test comparability and stability. As such, all stakeholders could profit from it, that is, laboratories, manufacturers, physicians, society, and naturally the patient. The medical laboratory, in particular, may be empowered for future tasks, such as contribution to the development, implementation, and management of global health-care policies.

References

1. Kayyali B, Knott D, Van Kuiken S. The big-data revolution in US health care: accelerating value and innovation. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care. (accessed December 2014)
2. Mayer-Schönberger V, Cukier K, editors. Big Data: A Revolution That Will Transform How We Live, Work, and Think. 3rd Ed. London: Houghton Mifflin Harcourt; 2013:242 p.
3. Williams NB. The Rise of Big Data, Trends and Opportunities for the Lab. Clin Lab News 2014;40(3):6-7.
4. Health Information Technology for Economic and Clinical Health (HITECH) Act, as enacted by the American Recovery and Reinvestment Act of 2009. Pub. L. 111-5, 2009
5. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule (45 CFR 170, Stage 1). Federal Register 2010;75(144):44590-654. http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17210.pdf
6. Ibid. Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; (45 CFR 170, Stage 2). Federal register 2012;77(171):54163-292. http://www.gpo.gov/fdsys/pkg/FR-2012-09-04/pdf/2012-20982.pdf
7. Ibid. Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Final Rule (42 CFR parts 412, 413, 422 et al, stage 1). Federal Register 2010;75(144):44314-588.

http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17207.pdf

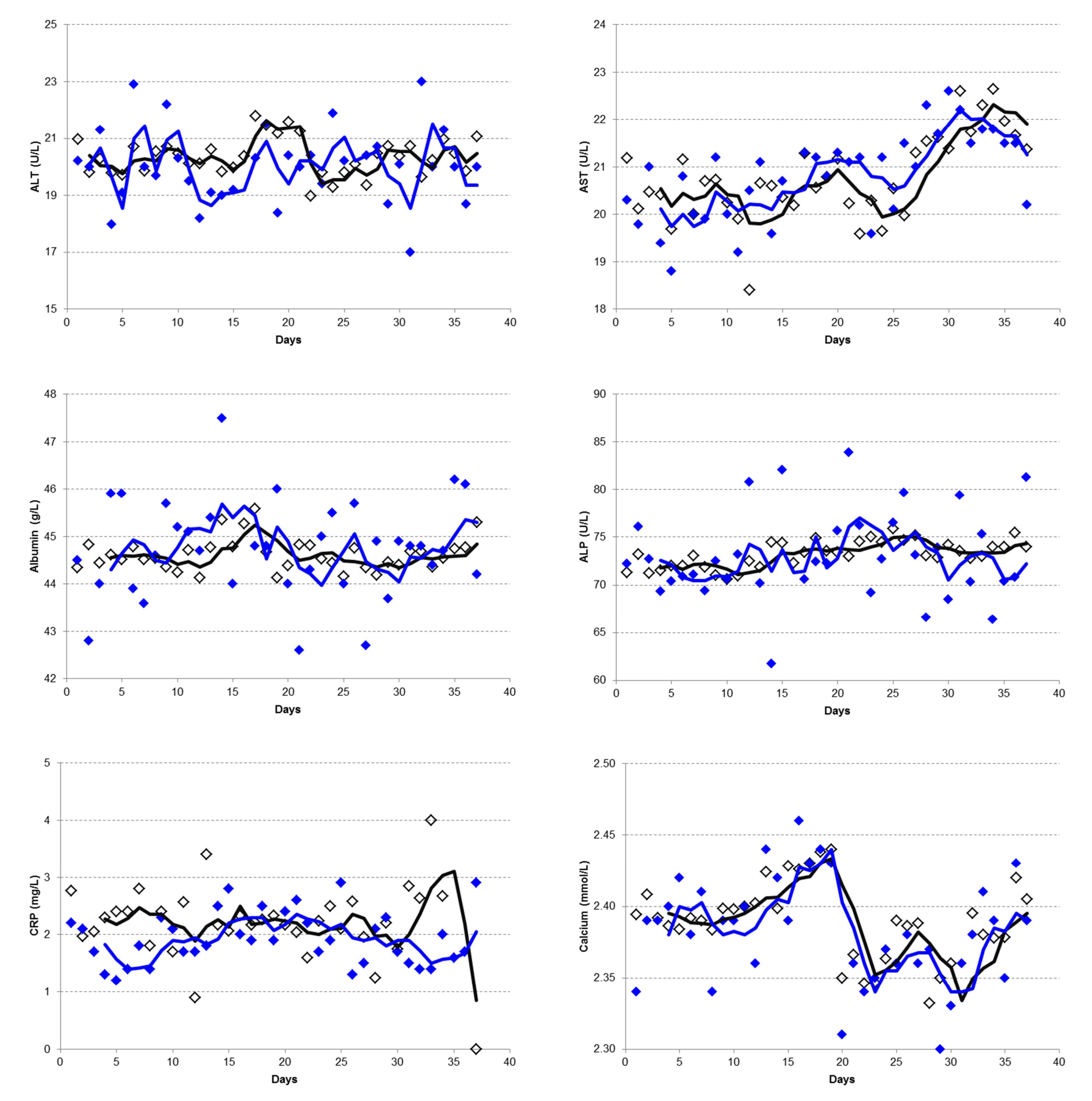
8. Ibid. Medicare and Medicaid Programs; Electronic Health Record Incentive Program - Stage 2 (42 CFR parts 412, 413, and 495). Federal Register 2012; 77(171):53968-4162. http://www.gpo.gov/fdsys/pkg/FR-2012-09-04/pdf/2012-21050.pdf
9. Frost & Sullivan, U.S. Hospital Health Data Analytics Market - Growing EHR Adoption Fuels a New Era in Analytics. http://www.frost.com/c/10046/sublib/display-report.do?id=NA03-01-00-00-00 (accessed December 2014).
10. O'Harrow R. The Machinery Behind Health-Care Reform. Washington Post 2009. http://www.washingtonpost.com/wp-dyn/content/article/2009/05/15/AR2009051503667.html (accessed December 2014).
11. Institute of Medicine. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington D.C.: The National Academies Press; 2012:212 p.
12. Van Houcke SK, Rustad P, Stepman HC, Kristensen GB, Stöckl D, Røraas TH, Sandberg S, Thienpont LM. Calcium, magnesium, albumin, and total protein measurement in serum as assessed with 20 fresh-frozen single-donation sera. Clin Chem 2012;58:1597-9.
13. Stepman HC, Tiikkainen U, Stöckl D, Vesper HW, Edwards SH, Laitinen H, et al. Measurements for 8 common analytes in native sera identify inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855-63.
14. Horowitz GL. Assessing accuracy on the front lines: a pragmatic approach for single-donor proficiency testing. Clin Chem 2014;60:806-8.
15. De Grande L, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont L. The Empower Project – A new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015 doi:10.1515/cclm-2014-0959.
16. Arbor Research Collaborative for Health. Dialysis Outcomes and Practice Patterns Study. http://www.dopps.org (accessed December 2014).
17. Solinger AB, Rothman SI. Risk of mortality associated with common laboratory tests: a novel, simple and meaningful way to set decision limits from data available in the Electronic Medical Record. Clin Chem Lab Med 2013;51:1803-13.
18. Westgard QC. Biological variation database, and quality specifications for imprecision, bias and total error (desirable and minimum). The 2014 update. http://www.westgard.com/biodatabase-2014-update.htm (accessed December 2014).
19. Goossens K, Van Uytfanghe K, Thienpont L. Trueness and

comparability assessment of widely used assays for 5 common enzymes and 3 electrolytes. Clin Chim Acta 2015 doi:10.1016/j.cca.2015.01.009.

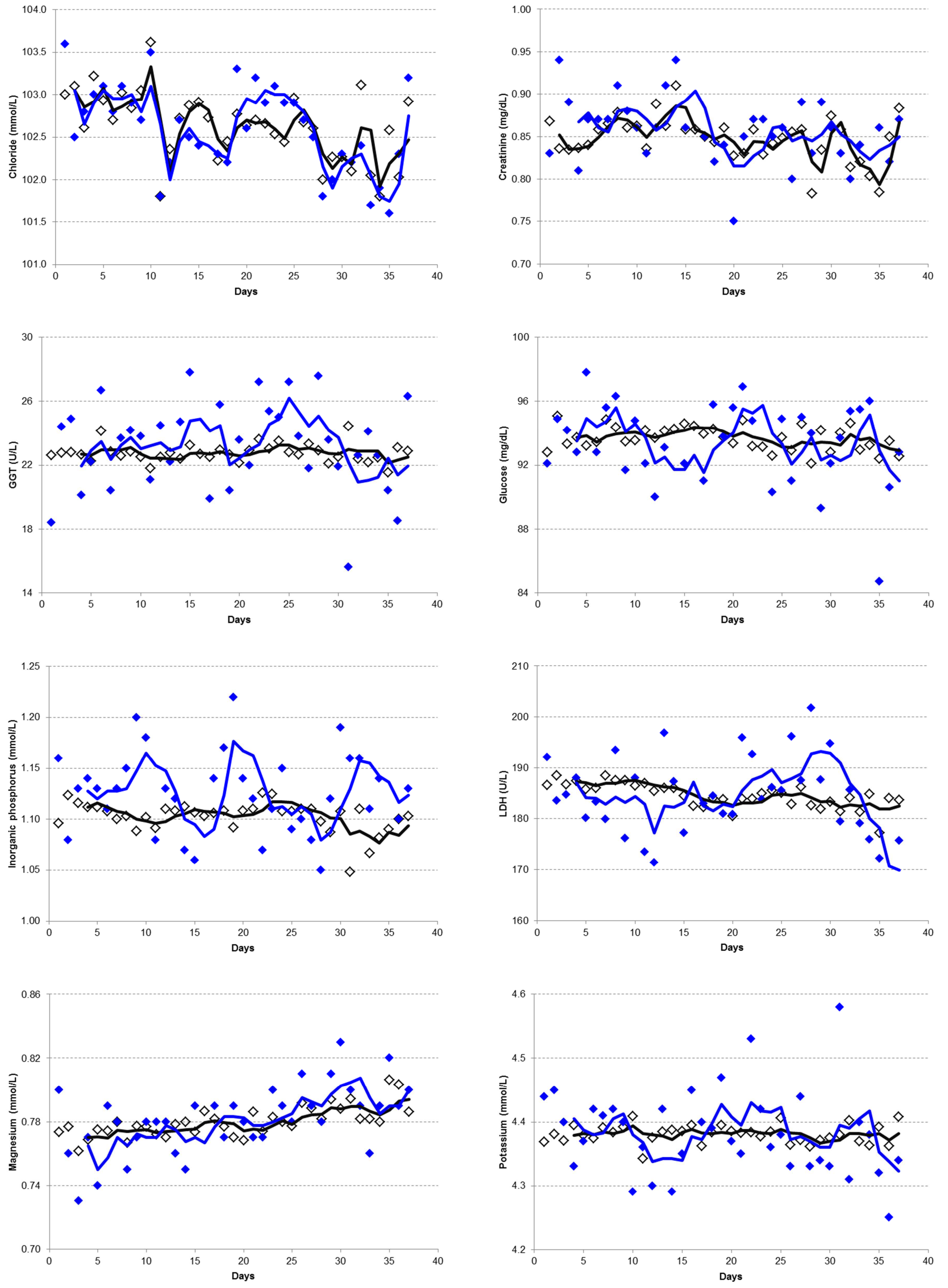
20. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
21. Braun SL, Ciolek K, Vogt W. Early detection of the selectivity loss of the Cobas Integra analyzer chloride ion-selective electrode. Clin Chem 2004;50:1272-3.
22. Stepman HC, Stöckl D, Stove V, Fiers T, Couck P, Gorus F, Thienpont LM. Long-term stability of clinical laboratory data: sodium as benchmark. Clin Chem 2011;57:1616-7.
23. Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P, et al. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.
24. Fleming JK, Katayev A. Changing the Paradigm of Laboratory Quality Control Through Implementation of Real-time Test Results Monitoring: For Patients by Patients. Clin Biochem 2014 Dec 27. pii: S0009-9120(14)00816-9. doi: 10.1016/j.clinbiochem.2014.12.016. [Epub ahead of print].
25. Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76-83.

Annex to Chapter 2

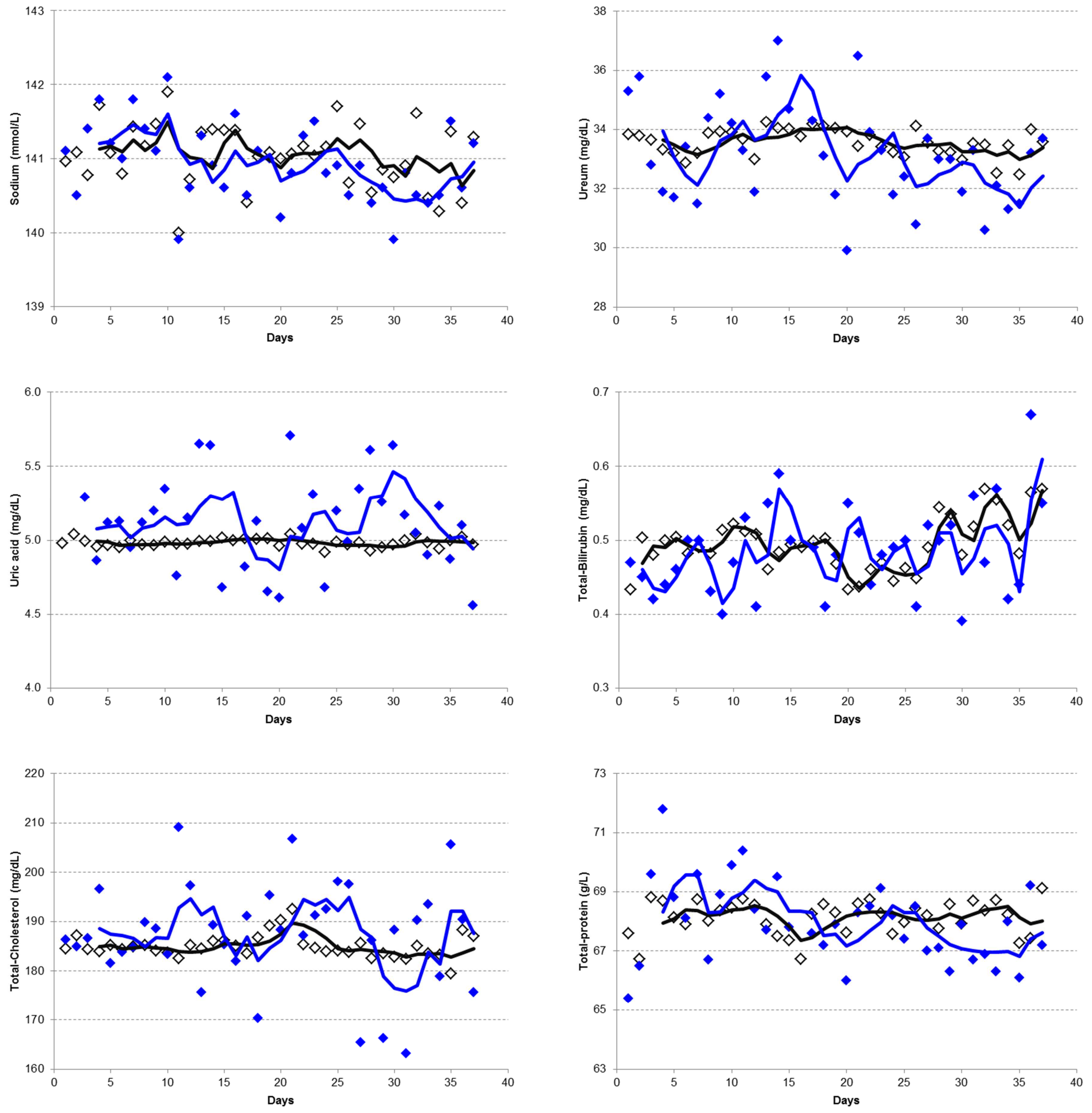
Comparison of IQC and Percentiler data



Supplemental figure 1A. Comparison of IQC (black) and Percentiler data (blue) for ALT, AST, albumin, ALP, CRP, and calcium.



Supplemental figure 1B. Comparison of IQC (black) and Percentiler data (blue) for chloride, creatinine, GGT, glucose, inorganic phospor, LDH, magnesium, and potassium.



Supplemental figure 1C. Comparison of IQC (black) and Percentiler data (blue) for sodium, ureum, uric acid, total-bilirubin, total-cholesterol, and total-protein.

3 Chapter III

The Flagger

This chapter is based upon:

- Goossens K, Brinkmann T, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Clin Chem Lab Med 2015;53:e269-70.
- Goossens K, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase (manuscript in preparation).
- Goossens K, De Grande L, Stöckl D, Van Uytfanghe K, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Euromedlab – 21st IFCC – EFML European congress of clinical chemistry and laboratory medicine (22-24 June 2014), Paris, France & 2015 AACC Annual Meeting and Clinical Lab Expo (26-30 July 2015), Atlanta, US [poster].

Highlights

The Percentiler has the potential of visualizing the effects of analytical instability directly on the results for patient samples, but it can’t show what effect instability has on the clinical outcome. Making the link between analytical variation and clinical outcome is very difficult, therefore we looked for an alternative in the form of the Flagger. This tool is capable of visualizing the instabilities of the hypo- and hyper-flagging rates, which we also refer to as “surrogate clinical outcome”.

Highlights of the research:

- An automatic platform for monitoring hypo- and hyper-flagging rates (“The Flagger”) was built.
- The Flagger data provide global evidence about test stability at low and high concentration ends.
- Both individual laboratories and manufacturers can use the Flagger to assess their problem analytes.
- The Flagger is capable of showing what stability can be achieved by current-state-of-the-art performance at low and high concentration ends.
- Combining data from the Flagger with the Percentiler has the potential to translate the effect of analytical variation on the flagging rate.

1. Introduction

With the Percentiler, participants are capable of assessing the stability of their instruments versus the limits we set. The latter are guided by the systematic error limits based on biological variation ("desirable" values from the Westgard website; 1), but they also take the current capability of diagnostic assays into account. When these limits are violated, however, it is difficult to define the clinical outcome of this aberration, because the link between analytical variation and clinical outcome is often unclear (2). That's why outcome-based analytical performance specifications, which reflect the clinical need of patients are so important. These specifications are at the top of the hierarchy of analytical performance limits as decided at the Stockholm conference of 1999 (3) and confirmed in the revision document written after the Milano congress in 2014 (4). These types of specifications are most valuable, but they are also difficult to calculate because testing is only indirectly linked to the clinical outcome of the test (2). To avoid the need of performing complex and expensive outcome studies, we will discuss a more practical alternative to translate the impact of analytical stability on laboratory medicine practice.

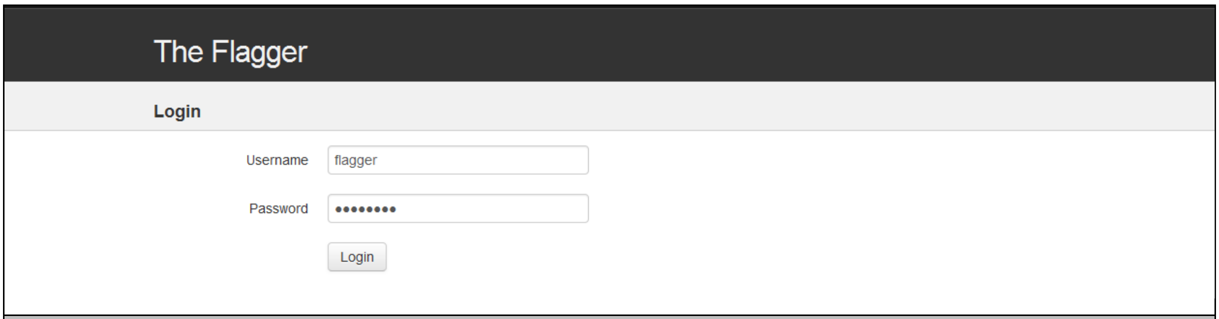
The first information of quantitative tests is usually expressed as a flagging outcome. When the value of a test result is located within the borders of a reference interval, the value is not flagged and not considered "suspicious". When it falls outside the reference interval, either lower or higher, the value is flagged. Although the outcome of the test isn't directly linked with analytical performance, the flagging outcome, or better said the flagging frequency of an assay, is. In fact, if one starts from the easiest model based on normal distribution, one is capable of calculating the effect of analytical variation or bias on the flagging rate. If one expresses those analytical changes relative to the biological variation of an analyte, one can determine, for example, theoretical limits which allow for a maximum relative change in the false discovery rate of 30%, as is described in the research of Stepman *et al* (5). But this is a theoretical model, and if one wants to learn what the real effect of analytical changes on the flagging rate is, one needs to establish a big database.

Therefore, we developed an online tool for monitoring flagging rates, which is called the Flagger. The tool makes use of the fact that the LIS alerts clinical biochemists about values which exceed locally chosen cut-offs. Similar to the Percentiler, the Flagger is a freely available online web tool, which receives instrument-specific, daily outpatient hypo- and hyper-flagging rates. The latter are calculated and transmitted by a laboratory from their middleware or LIS. The data are transmitted by e-mail, which is automatically read by software into a MySQL database. In the pilot phase the project focuses on the same analytes as in the Percentiler.

2. Flagger functionality

The Flagger functionality is analogous to the Percentiler functionality (pages 60-66). Only the differences between the two applications are discussed here. Just like the Percentiler, the Flagger application has been created by IT engineer Bruno Neckebroek.

Login



The Flagger

Login

Username flagger

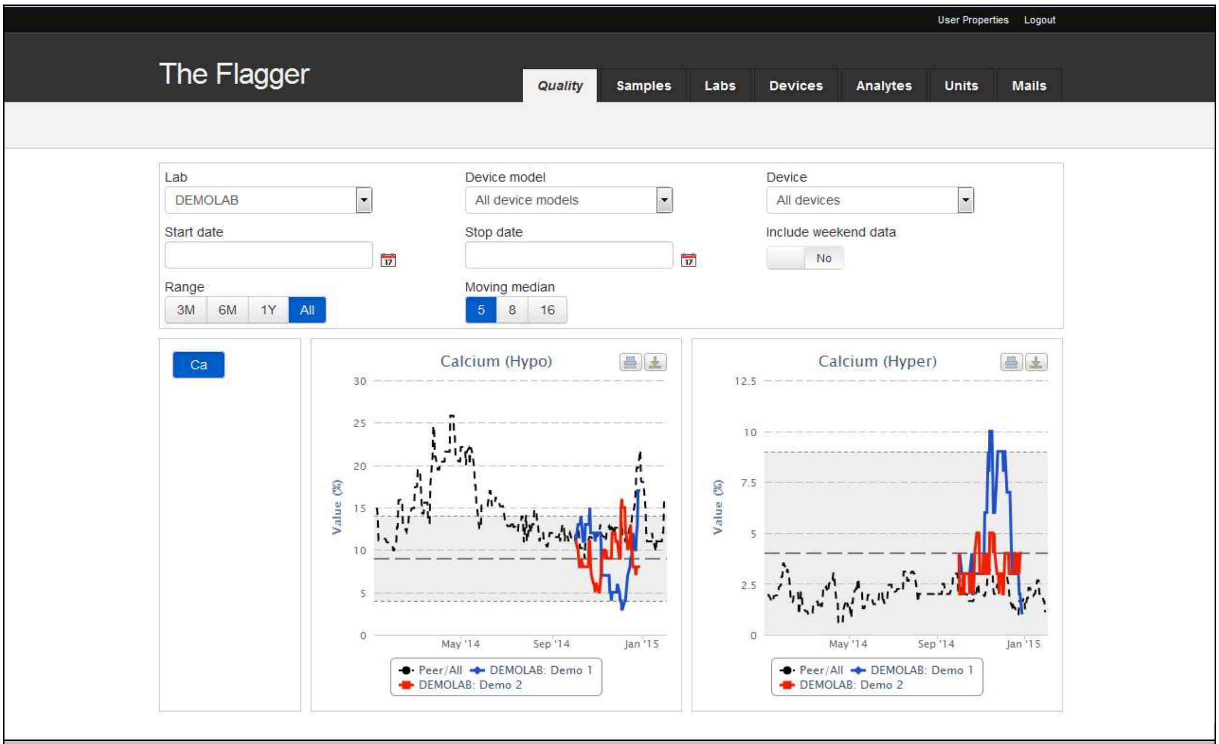
Password ••••••••

Figure 1. The Flagger login screen.

The Flagger is accessible online on: https://www.theflagger.be/. Usernames (not case sensitive) and passwords (case sensitive) are available for individual labs, for groups (e.g. organizations with multiple sites), and for the entire database (see Figure 1). Those who do not yet participate can make use of a demo-account (Username: DEMOLAB; Password: demo1234). Note that the data for the demonstration laboratory is only available for a limited number of analytes and for a short period in time. User guides are not available for the Flagger because this functionality mostly resembles the Percentiler.

The "Quality" tab

The selections which are made in the Quality tab, i.e., the laboratory in the "Lab" box, the peer group in the "Device model" box, the data range, and the *n* for the moving median, are structured in the same manner as in the Percentiler. However, after the selection, not one, but two charts will appear, without a summary table. The charts (example in Figure 2) show for the selected lab i) the moving median of hypo- (left graph) and hyper-percentage (right graph) of the selected devices over the selected period (full colored lines); ii) the long-term median (long-broken grey line) and the limits for the respective analyte (short-broken grey lines; the area in-between is shaded); iii) the peer group moving median (black broken line). It is possible to print and download the charts by pressing the respective buttons.



Lab: DEMOLAB

Device model: All device models

Device: All devices

Start date:

Stop date:

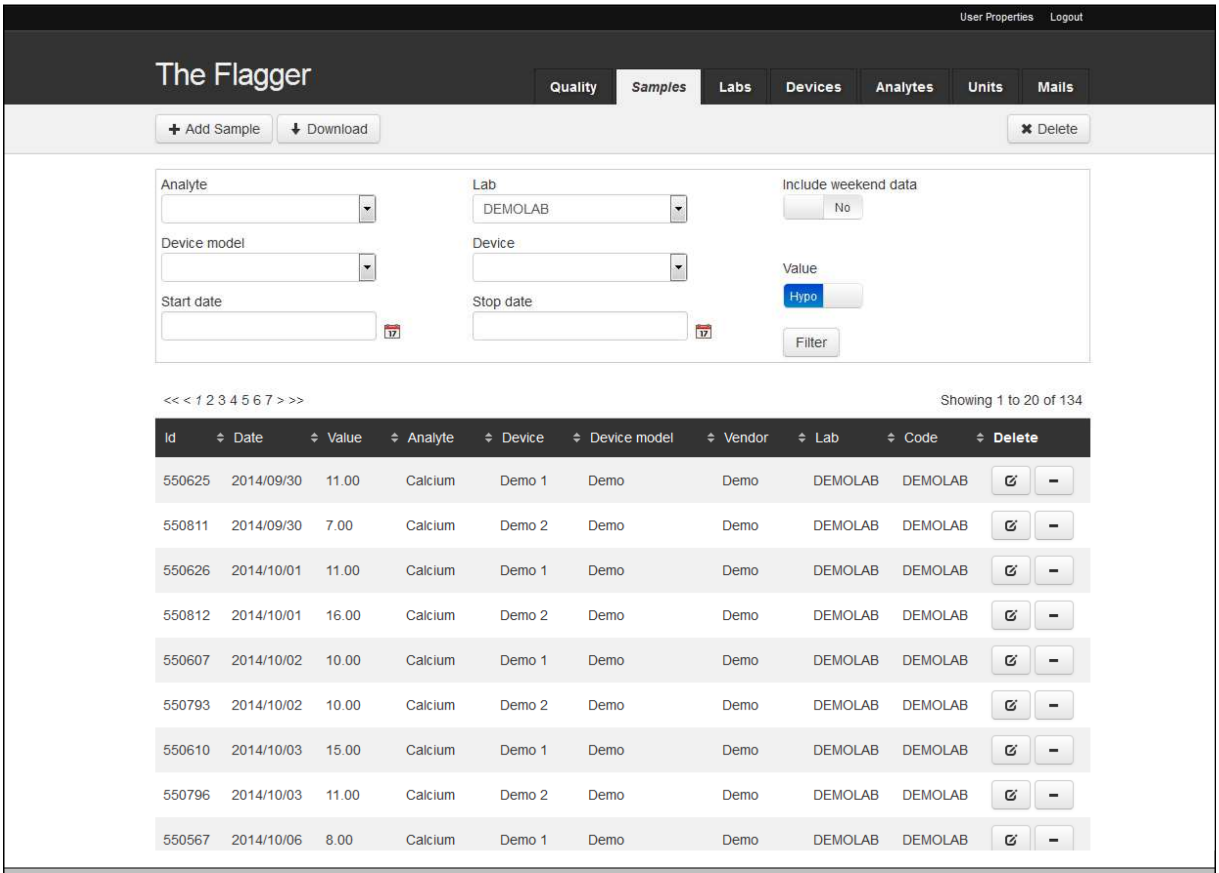
Include weekend data: No

Range: 3M 6M 1Y [✓ All]

Moving median: [✓ 5] 8 16

Figure 2. Demonstration of "The Flagger", enabling a laboratory to track the moving median of hypo- (left graph) and hyper-percentage (right graph) per instrument over time (Quality tab). The chart shows for the selected lab i) the moving median of the selected devices over the selected period (full colored lines); ii) the long-term median (long-broken grey line; represents the median calculated from all daily hypo- and hyper-flagging rates provided by the laboratory to which the graph applies) and the limits for the respective analyte (short-broken grey lines; the area in-between is shaded); iii) the peer group moving median (black broken line).

The “Samples” tab



Analyte:

Lab: DEMOLAB

Include weekend data: No

Device model:

Device:

Value: Hypo

Start date:

Stop date:

Showing 1 to 20 of 134

Id	Date	Value	Analyte	Device	Device model	Vendor	Lab	Code	Delete
550625	2014/09/30	11.00	Calcium	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550811	2014/09/30	7.00	Calcium	Demo 2	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550626	2014/10/01	11.00	Calcium	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550812	2014/10/01	16.00	Calcium	Demo 2	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550607	2014/10/02	10.00	Calcium	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550793	2014/10/02	10.00	Calcium	Demo 2	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550610	2014/10/03	15.00	Calcium	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550796	2014/10/03	11.00	Calcium	Demo 2	Demo	Demo	DEMOLAB	DEMOLAB	✎ −
550567	2014/10/06	8.00	Calcium	Demo 1	Demo	Demo	DEMOLAB	DEMOLAB	✎ −

Figure 3. Demonstration of “The Flagger” (Samples tab). The table shows 9 attributes for the selected lab: Lab identification number; date; median value (for hypo- or hyper-flagging rate); analyte; device identification; vendor or manufacturer; laboratory name; and lab ID code.

This window shows all entries in the database of the chosen laboratory for either the hypo- or hyper-flagging rate (see example Figure 3). Note, however, that data are mapped to the “Flagger” analyte names and units (not necessarily identical to the names and units given by the participants). One can filter the data (for example, according to analyte, to select in the “Analyte” box, always followed by pressing the “Filter” button) and sort (▲ ▼) (e.g. according to date, value, etc…). This tab also contains an excel exporting feature, where one can download the data, either completely or filtered. The other tabs available to the project team (Labs, Devices, Analytes, Units, and Mails) are similar as for the Percentiler; described in chapter 2.

Data transmission

Transition of the data is similar as for the Percentiler. Again, The data of each participant should be stratified for outpatients and ordered in a “table” consisting of data-related number of rows with ten fields. However, unlike for the Percentiler, data for the Flagger can be sent in two different formatting styles. The first possibility was already described in chapter 2 (page 64). The second option also makes use of ten fields which can be mentioned in any order (an example is shown in Figure 4):

1. Laboratory ID: chosen by the laboratory or already assigned by the laboratory
2. Date
3. Instrument ID: as already assigned by the laboratory
4. Outpatient code: three letters (e.g. OUT) or as already assigned by the laboratory
5. Analyte name: as already assigned by the laboratory
6. Unit: as already assigned by the laboratory
7. % flagged hypo
8. % “healthy”
9. % flagged hyper
10. Number of results used for calculating the median

Data can be transmitted as: i) e-mail embedded table; ii) e-mail attached EXCEL-file; or iii) e-mail attached text-file. Data are sent to a database-specific email address (flagger@stt-consulting.com) and is automatically read into a MySQL database.

e-mail embedded Table
From: ***
Sent: Saturday, 21 september 2015 06:31
To: flagger@stt-consulting.com
Content: Empower Flagger Project

Time produced : 27-09-2013 00:00 - 27-09-2013 23:59

```
ABCDEF;27/09/2013;C16000-5;POL;NA;mmol/L;140.9
ABCDEF;08/04/2015;COBAS8000C1;POL;ALB;g/L;        -   86%   14%    35
ABCDEF;08/04/2015;COBAS8000C2;POL; ALB;g/L;       -   87%   15%    37
ABCDEF;08/04/2015;COBAS8000C1;POL;ALT;U/L;        -   90%   10%   126
ABCDEF;08/04/2015;COBAS8000C2;POL; ALT;U/L;       -   72%   28%    57
ABCDEF;08/04/2015;COBAS8000C1;POL; CA;mmol/L;    4%   96%    -    119
ABCDEF;08/04/2015;COBAS8000C2;POL;CA;mmol/L;     6%   94%    -    102
ABCDEF;08/04/2015;COBAS8000C1;POL; CL;mmol/L;   12%   73%   15%    45
ABCDEF;08/04/2015;COBAS8000C2;POL;CL;mmol/L;    15%   83%    2%   104
ABCDEF;08/04/2015;COBAS8000C1;POL; GGT;U/L;       -   75%   25%   130
ABCDEF;08/04/2015;COBAS8000C2;POL;GGT;U/L;       9%   77%   14%   108
```

Figure 4. Example of data transition to the MySQL database of the Flagger.

3. Limits

Stability limits

Flagger limits, indicated with the horizontal, grey dotted lines in the graphs, are chosen so that they are in agreement with current state-of the-art performance at the higher and/or lower concentration ranges. Just like in the Percentiler, they are assigned semi-arbitrarily for each analyte (by spending a lot of time interpreting the Flagger data) so most of the participating laboratories belonging to the best performing peer group are able of achieving the set specifications. These limits are calculated relatively to the long-term flagging rate, but set an absolute minimum when the long-term flagging rate is low. For example, the relative limit for AST equals 30% with an absolute minimum of 1%. This means when a laboratory's long-term flagging rate equals 10%, the limits will be ± 3% (= 30% of 10%). When the long-term flagging rate equals 2.5% the limit equals the minimal 1%, and not 0.75%. An overview on the Flagger limits is provided in Table 1 of Article 3. This table shows that state-of-the-art performance for several analytes is capable of restricting the change of flagging rates to a maximum of 30% (relative compared to the long-term flagging rate). Even analytes with small biological variation, like potassium, can achieve this stringent quality specification. However, this table also indicates that current state-of-the-art performance at the low and high concentration levels requires increased stability especially for albumin, calcium, chloride, magnesium, sodium, inorganic phosphor, total-protein and uric acid (Flagging limit ≥ 50%).

Note, similar to the Percentiler the Flagger limits are dynamic, meaning they are adapted according to changes in state-of-the-art performance. Therefore, some Flagger figures in this thesis might still show a grey stability zone flanked by limits representing former numbers. Whenever the limits are functional to the provided explanation, however, the values and figures for the limits are adapted to the most recent numbers.

Target values

Typically, the reference interval used by a laboratory for a certain analyte is defined by 95% of the values in apparently healthy people. Theoretically, flagging rates of about 2.5% for both hypo-, and hyper-conditions are expected. However, for a number of analytes the normal distribution is skewed, e.g. for enzymes to higher concentrations (more hyper values). For other analytes general recommendations are applied, e.g. for total-cholesterol laboratories apply age-dependent treatment limits provided by The National Cholesterol Education Program (NCEP). As a consequence, the expected flagging rate will differ for all analytes, and due to a lack of consensus on the applied reference intervals, a target value is not provided for the Flagger application.

4. Comparison with Percentiler data

The stability of the hypo- and hyper-flagging rate is closely correlated to the calibration of the IVD instruments at the concentration ends of the reference interval. Therefore, the primary objective of the Flagger is monitoring the stability at the lower and higher concentration ends. Whilst separate monitoring of either patient percentile or flagging is extremely suited to follow the stability of the individual instrument, combining the two tools has the additional potential of comparing the trends in patient medians with the trends in the hypo- and hyper flagging rate (discussed in the article on pages 121-123). In principle this link between the Percentiler and Flagger tool is quite clear. When the values for the patient medians increase, the hypo-flagging rate will drop, and the hyper-flagging rate will increase and vice versa. In Figure 5, the Percentiler graphs (part 1) show instability for a glucose (A) and an LDH assay (B). These observations are also visible in the low and high concentration ranges, resulting in a significant difference for both the hypo- and hyper-flagging rate (respectively part 2 and part 3 of Figure 5). Note, for panel B2 the hypo-flagging rate doesn't follow the trend from the Percentiler because the long-term flagging rate equals zero.

In addition, when the population variation is low (or when the number of daily samples measured on a single instrument is high), it becomes possible to translate even small effects of an analytical shift on the flagging frequency. In Figure 6, the patient medians for calcium shift to values that are ~0.03 mmol/L lower. Although this effect is considered insignificant, the Percentiler and Flagger graphs are still concordant. In this way, the laboratory is able to translate the effect of the shift in the Percentiler on the flagging frequency. The hypo-flagging rate increases, insignificantly, from ~3% to ~4%, and the hyper-flagging rate decreases, again insignificantly, from ~5% to ~3%.

Note, however, that the link between Percentiler and Flagger observations is not always confirmed. If a bias-related issue or another analytical problem occurs, which is restricted to a certain concentration level, then the Percentiler will not always confirm Flagger observation or vice versa. In Figure 7 the Percentiler shows a bias between two instruments which measure magnesium. The hyper-flagging rate is concordant with this bias observation, while the hypo-flagging rate is similar for both instruments. Although this might seem strange, this observation can be explained by a difference in calibration at different concentration ranges. This also indicates that several assays might be more stable at lower and/or higher concentration levels than they are at the median concentration, or vice versa.

An extensive discussion on the Percentiler-Flagger link, with an example for each analyte, can be found in the manuscript in preparation on pages 124-144.

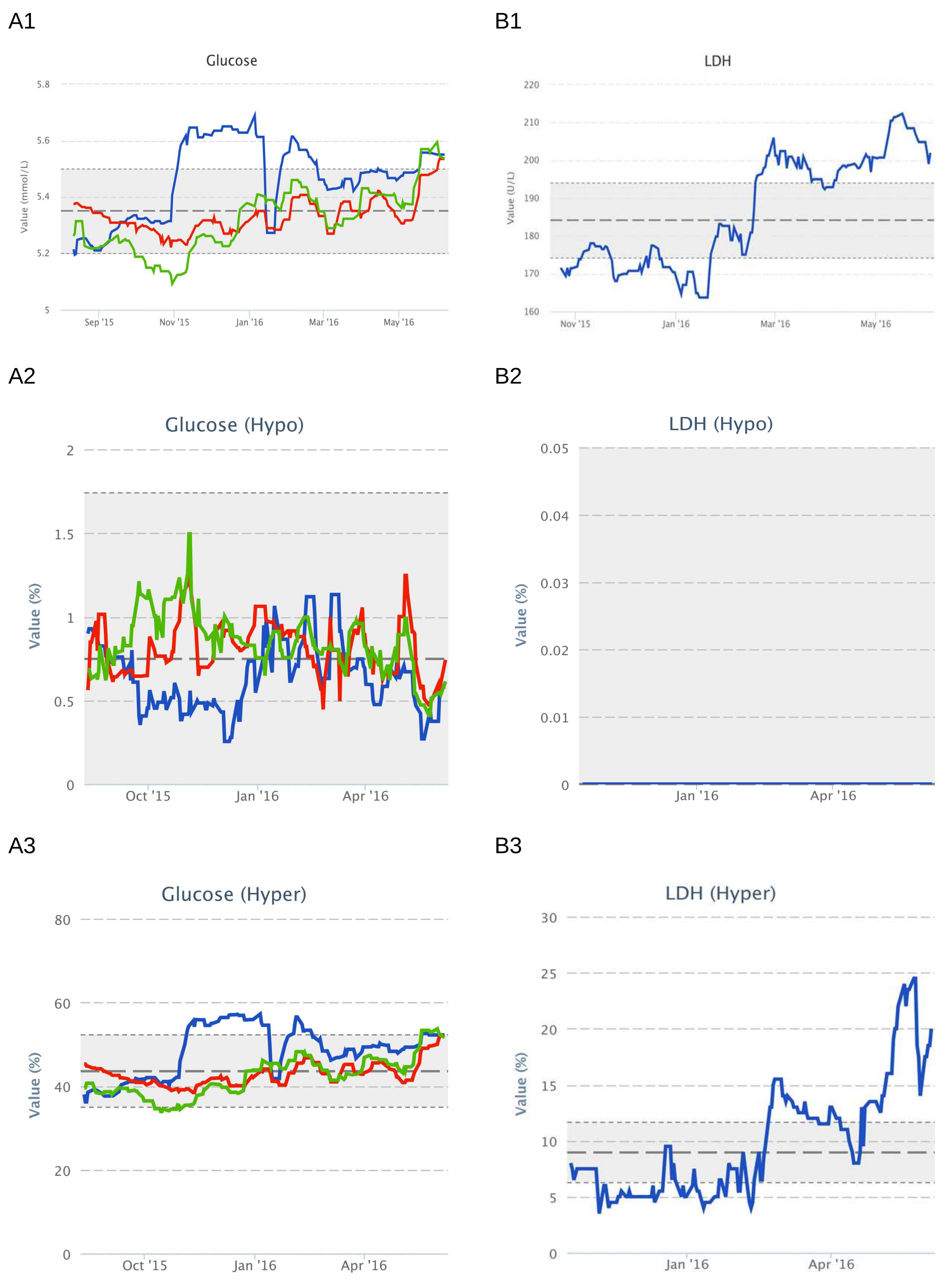


Figure 5. Percentiler graph showing trends for the moving median of patient results (1) for (A) glucose and (B) LDH, which is mirrored in the Flagger for both the hypo- and hyper-flagging rate (2 and 3 respectively).

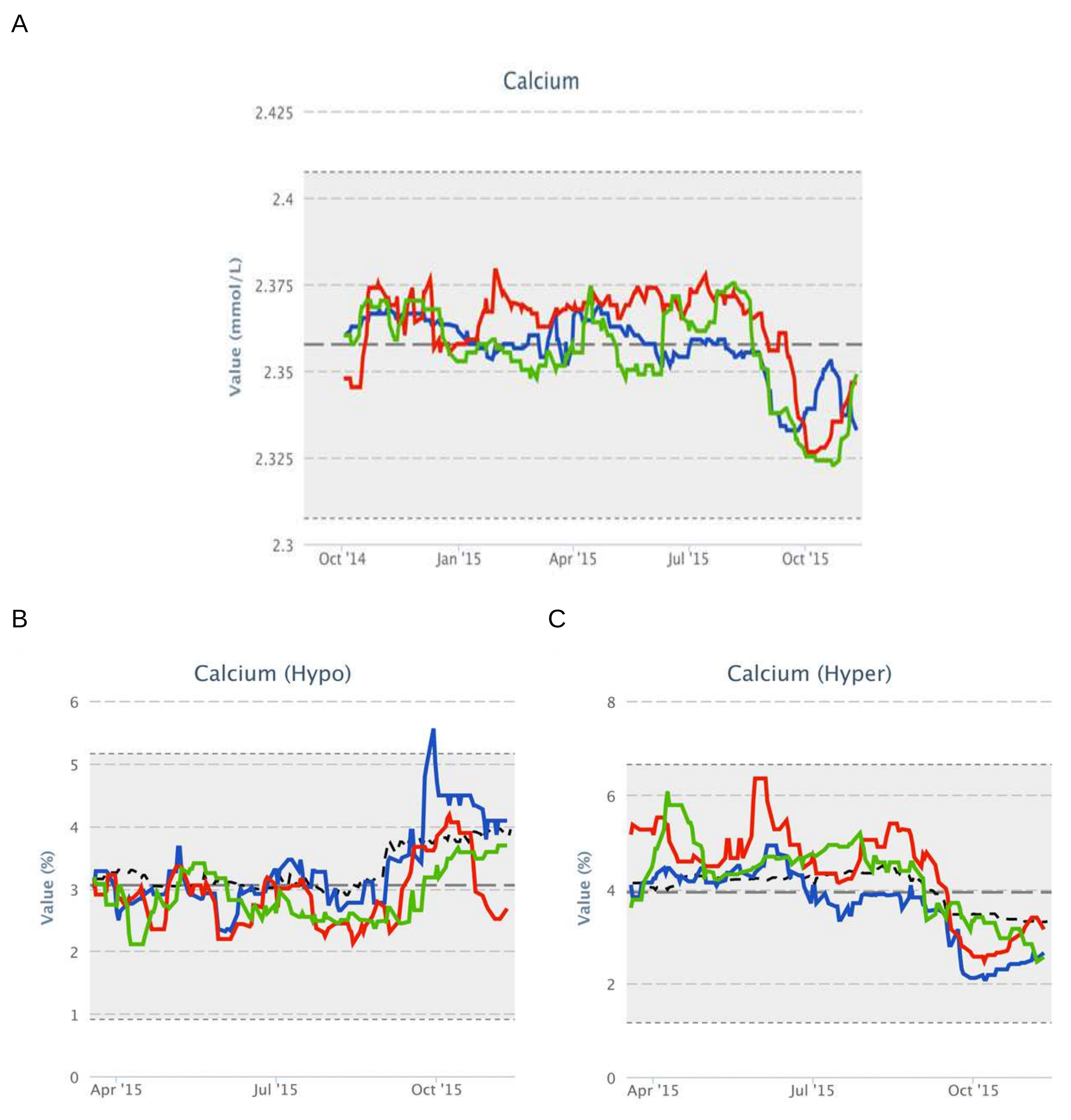


Figure 6. Percentiler graph showing trends for calcium: the patient medians for calcium shift to values which are ~0.03 mmol/L lower (A). This is mirrored in the Flagger: the hypo-flagging rate increases from ~3% to ~4% (B), and the hyper-flagging rate decreases from ~5% to ~3% (C).

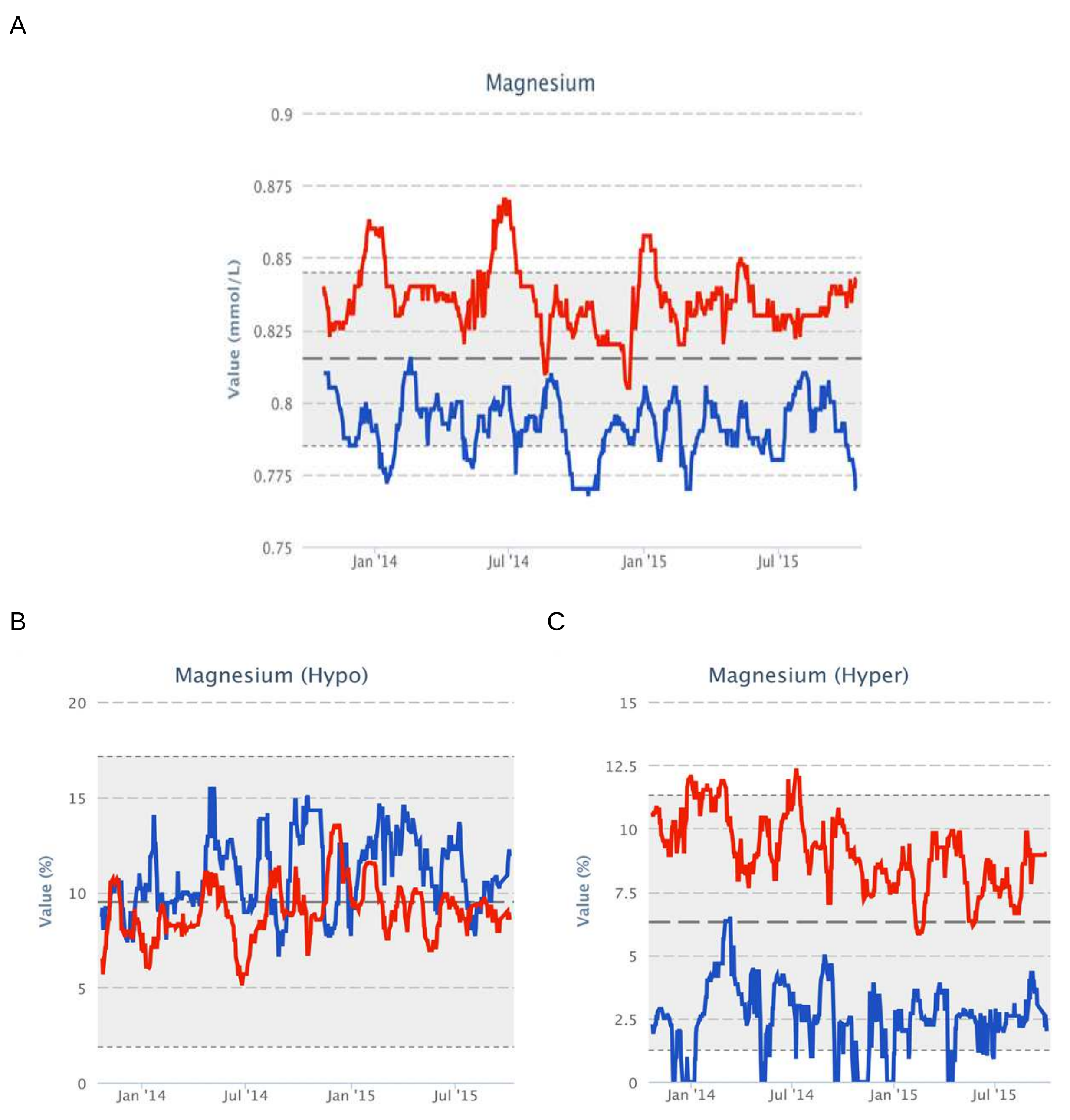


Figure 7. Percentiler graph showing a bias between two instruments measuring magnesium (A). The effect of the bias is translated in the hyper-flagging rate (C), but not in the hypo-flagging rate (B).

5. Conclusion on the Flagger

With our previously discussed Percentiler application we can show the currently achievable state-of-the-art performance. However, this tool cannot give an indication of the effect of violations of the stability limits on the clinical outcome or medical decision. Here we discussed a second application, called the Flagger which monitors the stability of the flagging rate in relation to the stability of the patient medians. Just like the Percentiler, the Flagger has several advantages and disadvantages over the existing QC tools. The information given by the Flagger also originates from materials which are "as commutable as possibly can be". They are again available without additional costs. Naturally, the Flagger's utility improves when the population variation of the data is low. In theory, participants can use the Flagger just like the Percentiler to monitor the stability of their IVD instruments. Combining the two tools has the additional potential of comparing the trends in patient medians with the trends in the hypo- and hyper flagging rate. A laboratory, thereby, still cannot translate the effect of violations of bias limits on the clinical outcome, but it can at least visualize the effect on the flagging rate.

References

1. Westgard QC. Desirable specifications for total error, imprecision, and bias, derived from biologic variation. https://www.westgard. com/biodatabase1.htm (accessed January 2014)
2. Horvath AR, Bossuyt PM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WD, Lennartz L, Cobbaert CM, Ebert C, Lord SJ; Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine. Setting analytical performance specifications based on outcome studies - is it possible? Clin Chem Lab Med 2015;53(6):841-8.
3. Kallner A, McQueen M, Heuck C. The Stockholm Consensus Conference on quality specifications in laboratory medicine, 25-26 April 1999. Scand J Clin Lab Invest 1999;59(7):475-585.
4. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, Petersen PH, Schimmel H, Sikaris K, Panteghini M. 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(6):833-5.
5. Stepman HC, Stöckl D, Twomey PJ, Thienpont LM. A fresh look at analytical performance specifications from biological variation. Clin Chim Acta 2013;421:191-2.

Article 2. The Flagger concept (letter to editor)

On-line flagging monitoring – A new quality management tool for the analytical phase

Kenneth Goossens[1], Thomas Brinkmann[2], Linda M Thienpont[1*]

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
[2]Labor Lademannbogen MVZ GmbH, Hamburg, Germany

Traditionally, it is difficult to demonstrate the impact of analytical quality on daily medical decision making (1). This is partly due to the fact that analytical quality should relate to its effect on clinical decisions (2). Strictly, this requires complex and expensive outcome studies, however, in the field of laboratory medicine these are still lacking. We, therefore, looked for alternative tools to translate the impact of analytical quality, particularly assay stability, on daily laboratory medicine practice. We conducted 2 pilot studies to investigate the utility of a tool demonstrating the effect of analytical shifts on so-called "surrogate" medical decisions, such as flagging of laboratory results exceeding locally used cut-offs (3, 4). On the basis of the observations made in these preliminary studies, we elaborated a theoretical concept (5). Because of the encouraging comments we received for that preliminary work, we developed an on-line tool for monitoring flagging rates, which we called "The Flagger" (www.theflagger.be; login: demolab, password: demo1234). The tool makes use of the fact that laboratory information systems (LIS) alert clinical biochemists about values that exceed locally chosen cut-offs, for example, by making them bold in the report or by adding an asterisk to the result (here referred to as "flagging"). The LIS is programmed for our application to calculate the daily %-flagging rates. Those numbers are automatically assembled, either in e-mail embedded lists, EXCEL-files, or text files. These files are automatically sent to our database and visualized by our Flagger application.

Here, we report our first experiences with "hypercalcemia" as example of surrogate medical decision (see Fig. 1). We used age-dependent flagging cut-offs: 2.6 mmol/L (10 days of age); 2.75 mmol/L (2 years), 2.7 mmol/L (12 years); 2.55 mmol/L (18 years); 2.5 mmol/L (60 years); 2.55 mmol/L (90 years); and 2.4 mmol/L (>90 years). The daily medians were calculated from ~75 outpatient results and the data had a CV of ~4%. The figure shows the time-course of the daily calcium medians (left graph) and the flagging frequency indicating “hypercalcemia” (right graph) for two instruments in a laboratory. Instrument B (red lines) was stable over the observed time-period, instrument A (blue lines) exhibited a shift of ~0.06 mmol/L (~2.5%). Typically, laboratories would not consider a shift of this magnitude of major concern. However, if they are alerted by a tool like the Flagger that this shift translates into a ~3-fold increase of the flagging rate (from ~3% to ~9%), they should be given an incentive to stricter quality control, in particular, for analytes that are under narrow physiological control (such as sodium, or calcium). In the given case, it may be desirable to keep the stability of the calcium assay within 0.04 mmol/L. Note that the aforementioned increase corresponds very well to the expected 3.5-fold increase when the shift is ~0.63 times the CV (5), as is in our case (2.5% shift/4.0% CV = 0.63). We would like to emphasize that the Flagger can be used “generic”, meaning that the bias/population CV ratio can predict to a good extent the expected change in flagging rate also for other analytes.

The Flagger is an interesting new quality management tool because it directly translates analytical shifts into surrogate medical decision. The actual utility of the Flagger and its impact on medical decisions, naturally, will have to be inferred from practice. However, in view of the experience that requestors of laboratory analyses such as general practitioners use to highly rely on flagging of the reported results, it can be expected that an “analytically stable” flagging rate

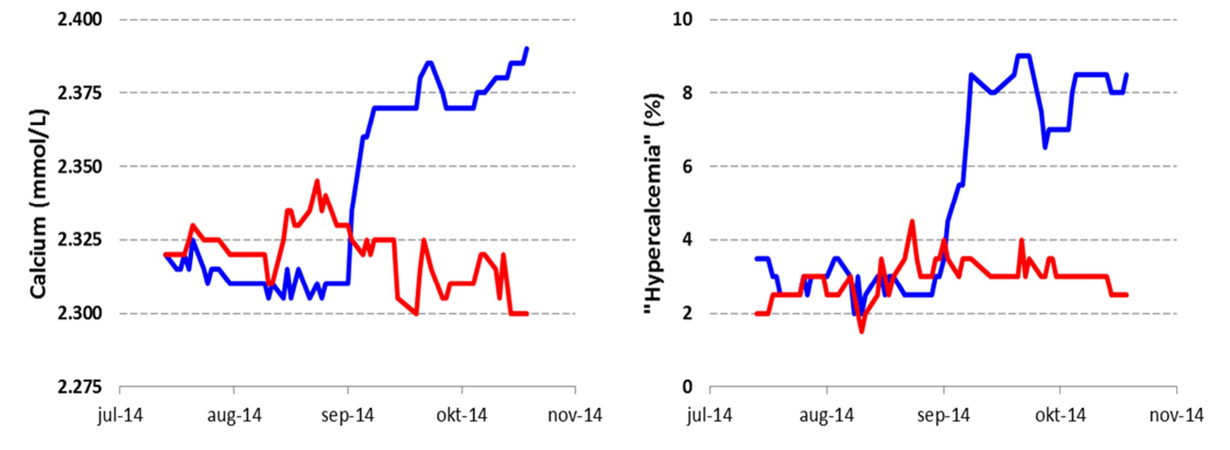


Figure 1. Moving medians for calcium in time calculated from 8 daily outpatient medians (left graph) and associated flagging percentage indicating “hypercalcemia” (right graph) for two instruments (A: blue lines; B: red lines) used in a laboratory; the mmol/L axis ranges over ~5% (2.275 – 2.4 mmol/L).

consequently, to related medical decisions. Currently, this tool is still to be programmed by the local IT-department, however, we found laboratory information system providers interested to develop generally applicable solutions. Moreover, with our Flagger platform, we are able to peer group monitoring of flagging rates opening all the benefits of peer group comparisons.

In conclusion, we consider on-line monitoring of the flagging rate in the individual laboratory, but also externally at the peer group level, an additional quality management tool for the analytical phase. It is particularly useful because it directly translates analytical quality into quality of medical decision making against locally important cut-offs.

References

1. Stöckl, D. Modern quality management misunderstood? Clin Chem 1998;44:1066-7.
2. Fraser CG, Kallner A, Kenny D, Petersen PH. Introduction: strategies to set global quality specifications in laboratory medicine. Scand J Clin Lab Invest 1999;59:477-8.
3. Stepman HC, Stöckl D, Stove V, Fiers T, Couck P, Gorus F, Thienpont LM. Long-term stability of clinical laboratory data: sodium as benchmark. Clin Chem 2011;57:1616-7.
4. Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P, Anckaert E, Gorus F. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.
5. Stepman HC, Stöckl D, Twomey PJ, Thienpont LM. A fresh look at analytical performance specifications from biological variation. Clin Chim Acta 2013;421:191-2.

Article 3. The Flagger (manuscript submitted to Clinica Chimica Acta)

The Flagger – Relating analytical performance to surrogate medical decisions

Kenneth Goossens[1], Linda M Thienpont[2*], and participating laboratories

[1]Department of Pharmaceutical Analysis, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

[2]Thienpont & Stöckl Wissenschaftliches Consulting GbR, Erlbacher Strasse 11, 86643 Rennertshofen (OT Bertoldsheim), Germany

Introduction

The laboratory medicine community is concerned about analytical performance, however, struggles since long with defining specifications to assess quality. First laboratory inter-comparisons were validated by use of arbitrary specifications inferred from expert advice (1). Later on, several scientific approaches were developed based on the biological variation of the analytes (2, 3). Those were followed by many other models, including specifications derived from questionnaires to clinicians and requirements for decisions in specific clinical situations (4). A milestone was the Stockholm conference, defining a five-level hierarchy for establishing analytical performance specifications, with specifications derived from "clinical situations" at the top (5). Unfortunately, little progress was made with implementing generally accepted "numbers", which gave rise to holding a conference that addressed the situation 15 years later (6, 7). In principle, the original proposal was re-iterated, however, the five-level hierarchy was streamlined into three levels with the clinically derived specifications at the top. Unfortunately, the latter give only very few generally accepted numbers for routine practice in the laboratory. Therefore, specifications based on biological variation are still the most widely used ones. Only, there is no consensus about the actually desired numbers (optimum, desirable, minimum) and numbers that really can be applied in practice. Because of that, we expanded the ideas of Klee (8), investigating analytical instabilities (in fractions of the biological variation) onto their effect on surrogate medical decisions (9). Based on that theoretical framework, we developed the Percentiler and Flagger applications that monitor analytical stability (Percentiler) and its effect on flagging rates (surrogate medical decisions) in the individual laboratory (10-16). Combining both applications has the chance to bridge the medium hierarchy level (biological variation) with the top hierarchy level (clinical outcome) and thus helping the individual laboratory in their local situation to define realistic, but ambitious analytical performance specifications.

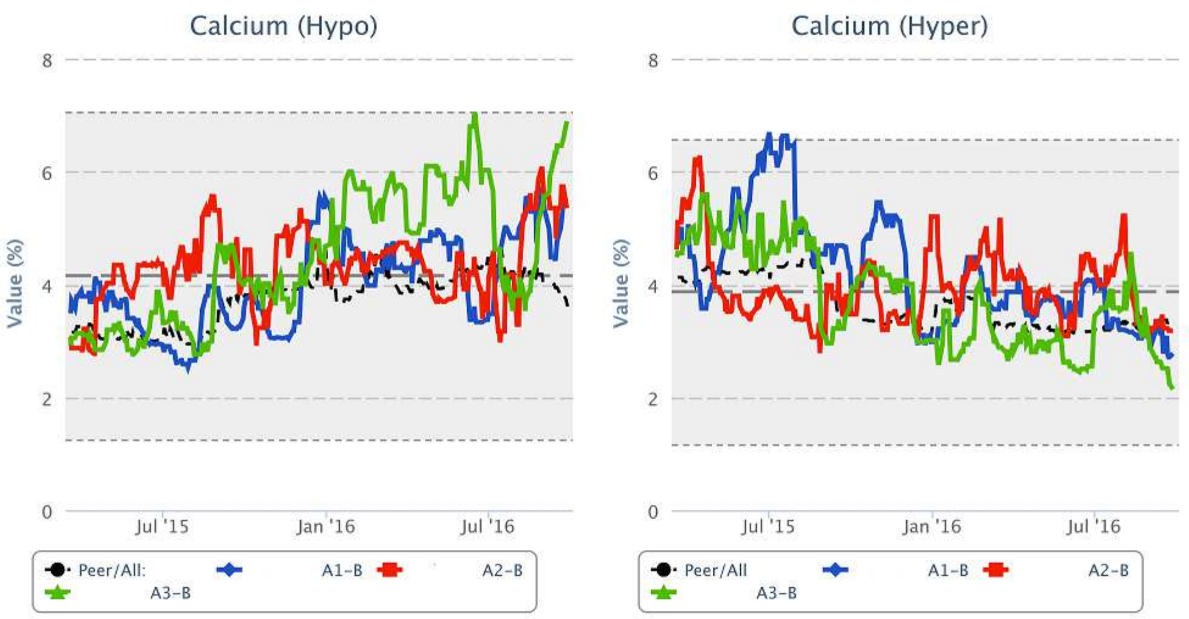
Here, we report our longer term experience with both applications and demonstrate the effects of analytical variation on the local hypo and hyper flagging rates with the aim of developing analytical performance specifications that can be applied in daily routine practice and different laboratory surroundings.

Materials and methods

The Flagger is a freely available online web tool that receives instrument-specific, daily outpatient hypo- and hyper-flagging rates (in percentage of the total number of daily results). The latter are calculated and transmitted by the participating laboratories from their middleware or laboratory information system. The data are sent by e-mail and automatically read by software into our MySQL database. In the pilot phase the project focuses on 20 analytes from clinical chemistry tests, i.e., albumin (ALB); alkaline phosphatase (ALP); alanine aminotransferase (ALT); aspartate aminotransferase (AST); total-bilirubin (BIL); calcium (CA); total-cholesterol (CHOL); chloride (Cl); C-reactive protein (CRP); γ-glutamyltransferase (GGT); glucose (GLU); potassium (K); creatinine (CRE); lactate dehydrogenase (LDH); magnesium (MG); sodium (NA); inorganic phosphor (PHOS); total-protein (PROT); urea or bound urea nitrogen (UREA; BUN); and uric acid (UAC). In a later stage, the database was extended to free thyroxine (FT4) and thyroid-stimulating hormone (TSH) in the context of the IFCC standardization/ harmonization of thyroid function tests (11).

The Flagger is accessible online on: https://www.theflagger.be/. Usernames and passwords are available for individual laboratories, for groups (e.g. organizations with multiple sites), and for the entire database. Those who do not yet participate can access the online web tool by use of a demo account (Username: DEMOLAB; Password: demo1234). Only the Empower team has access to the data of all laboratories, while participants are restricted to view their own data. Currently, 51 laboratories are participating in the Flagger application with 135 instruments. The longest participation time is 2.5 years. Cost-free automatic data sending by several Laboratory Information Systems is available since 1.5 years. All "Flagger" laboratories, also, participate in the Percentiler application (12).

The Flagger charts (example shown in the online Supplemental, Figure 1S) show for the selected laboratory (i) the moving median of the daily hypo- and hyper percentage, (ii) the long-term median and the limits for the respective analyte (relative to the long-term flagging median), and (iii) the all or peer group moving median. In the web application



Supplemental Figure 1S. The Flagger charts show (i) the moving median of daily hypo- and hyper-percentage of the selected devices over the selected period (full colored lines); (ii) the long-term median (long-broken grey line) and the limits for the respective analyte (short-broken grey lines; the area in-between is shaded); (iii) the all or peer group moving median (black broken line).

several selections (via buttons) are possible. One can choose the device type, in- or exclude weekend data and the time frame. The latter can be done by pushing the buttons “Start date”, “Stop date” or “3M” (months), “6M”, “1Y” (year), and “All”. One also needs to choose the n for calculating the moving median (default setting is 5), and finally the analyte for which the chart should be shown. For more information on the Percentiler functionalities we refer to our previous publication (12).

Results

The combination of the Flagger and Percentiler application allowed deriving performance specifications for hypo and hyper flagging rates and their comparison with the respective analytical performance specifications for 22 commonly measured analytes (see Table 1). The limits were achievable by at least one quarter of the laboratories. The most stringent ones (20%) were derived for CHOL and GLU because of their paramount importance for public health policies (coronary artery disease and diabetes). A Flagger limit of 30% could be applied for 11 analytes. Higher limits (50% and 70%) had to be used for the remaining 8 analytes, in particular, those with a quite small biological variation (CA, MG, NA).

Table 1. Limits for the Flagger and Percentiler applications and desirable bias from biological variation (17); #note: the minimum absolute %-limit in the flagger is 1%; table sorted by increasing Flagger limit.

Analyte	Flagger# limit (%)	Percentiler limit (%) (Absolute)	Desirable stability limit (%) (17)
CHOL	20	3.8 (0.2 mmol/L)	4.1
GLUC	20	3.1 (0.15 mmol/L)	2.3
ALKFOS	30	6.8 (5 U/L)	6.7
ALT	30	9.5 (2 U/L)	11.5
AST	30	6.5 (1.5 U/L)	6.5
BILTOT	30	10.0 (1 µmol/L)	9.0
CREAT	30	3.9 (3 µmol/L)	4.0
CRP	30	9.6 (0.25 mg/L)	21.8
FT4	30	3.3 (0.5 pmol/L)	3.3
GGT	30	9.1 (2 U/L)	11.1
K	30	2.4 (0.1 mmol/L)	1.8
LDH	30	4.6 (8 U/L)	4.3
TSH	30	7.7 (0.12 mU/L)	7.8
UREA	30	6.0 (0.3 mmol/L)	5.6
ALB	50	2.3 (1 g/L)	1.4
CL	50	1.0 (1 mmol/L)	0.5
PHOS	50	4.4 (0.05 mmol/L)	3.4
PROT	50	1.4 (1 g/L)	1.4
URIC ACID	50$	4.8 (15 µmol/L)	4.9
CA	70	1.7 (0.04 mmol/L)	0.8
MG	70	3.0 (0.02 mmol/L)	1.8
NA	70	0.7 (1 mmol/L)	0.2

$50% instead of 30% because of seasonal variation

The effect of analytical variation – and eventual violation of the Flagger limit – is exemplified for five selected examples (see the Supplemental Figures 2S to 22S for all 22 analytes).

Cholesterol

Figure 1 shows a major downwards shift in total-cholesterol median values from ~4.7 mmol/L to ~4.4 mmol/L. The hypo flagging rate is nearly unaffected and is in the order of ~1%, while the hyper flagging rate decreases from ~36% to nearly 23%. Operating the test within the ±0.2 mmol/L limit would result in hyper flagging rates between 24 and 37%.

Glucose

Figure 2 shows three shifts in glucose values, i.e., a first time from ~5.1 mmol/L to ~5.3 mmol/L, then a shortly lasting shift up to 5.5 mmol/L followed by a return to ~5.4 mmol/L, and finally a third shift up to ~5.6 mmol/L. The hypo flagging rate is little affected and is in the order of 1% to 2%, while the hyper flagging rate mainly increases due to the last upwards shift from ~35% to nearly 55%. Operating the test within the ±0.15 mmol/L limit would result in hyper flagging rates between 33 and 50%.

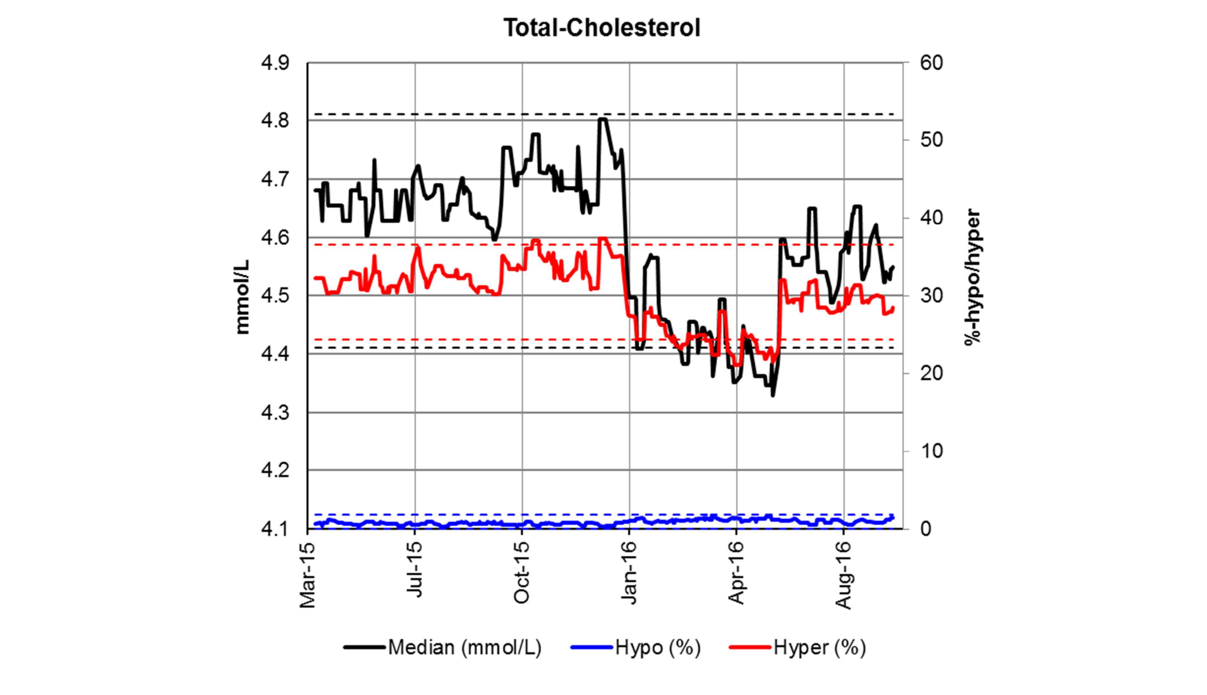


Figure 1. The figure shows the time courses of (i) the moving median of daily total-cholesterol values (full black line), (ii) the hypo flagging rate (full blue line), (iii) the hyper flagging rate (full red line), and (iv) the respective limits indicated by broken lines in the same color as the parent lines. Note, the Percentiler limit is 0.2 mmol/L (= 3.8%; “desirable” = 4.1%) and the Flagger limit is 20% of the long-term laboratory median.

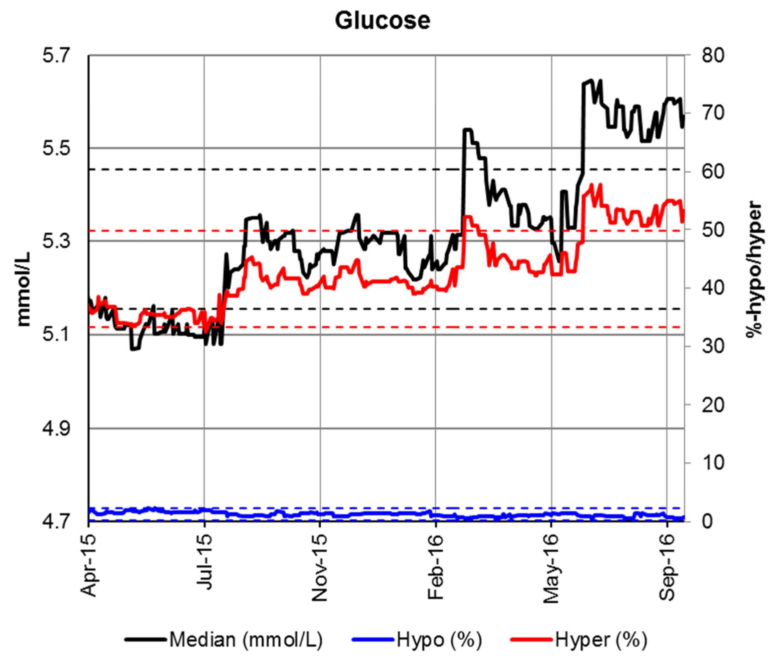


Figure 2. Similar to Figure 1, but for glucose. Note, the Percentiler limit is 0.15 mmol/L (= 3.1%; "desirable" = 2.3%) and the Flagger limit is 20% of the long-term laboratory median.

ALT

Figure 3 shows a drift of the ALT values from ~24 U/L to ~18 U/L and a sharp shift to ~30 U/L. The hypo flagging rate increases from ~10% to ~28% and drops to 0%. The hyper flagging rate decreases from ~7% to nearly 0% and increases, again, to ~8%. Operating the test within the ±2 U/L limit would result in hypo flagging rates between 9 and 17%. Note, typically, ALT tests were stable within the selected Flagger limit of 30% (data not shown). Note also that most laboratories have hyper flagging limits, only, for ALT and the other monitored enzymes.

Calcium

Figure 4 shows moderately varying calcium values, i.e., they first shift upwards from ~2.33 mmol/L to ~2.38 mmol/L, then they fall back to 2.33 mmol/L, followed by a second gradual increase to ~2.42 mmol/L, to finally drop back to 2.30 mmol/L. The hypo flagging rate decreases more strongly with the analytical shifts from ~13% to ~3%, then increases to ~9%, falls back to 4% to finally increase a second time to ~15%. The hyper flagging rate is very low so that no effects are observed. Operating the test within the ±0.04 mmol/L limit would result in hypo flagging rates between 2 and 10%.

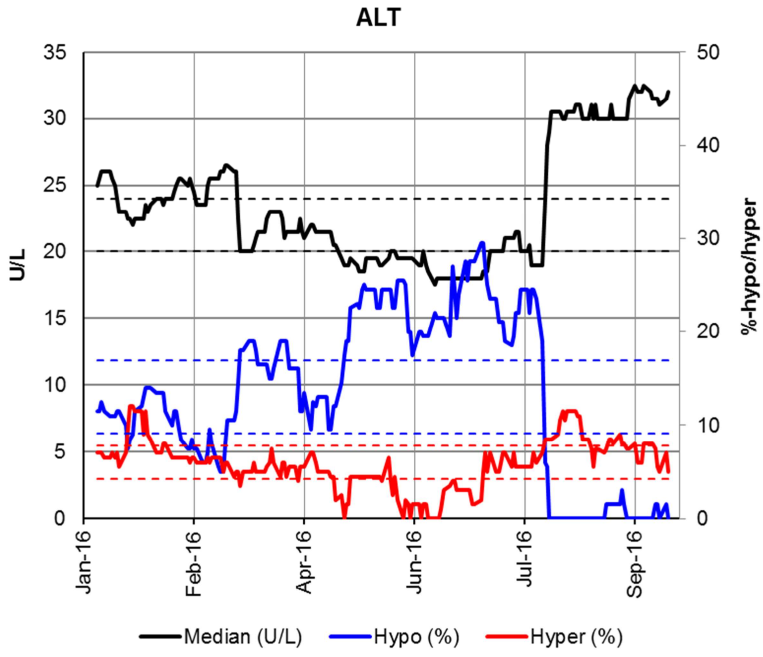


Figure 3. Similar to Figure 1, but for ALT. Note, the Percentiler limit is 2 U/L (= 9.5%; "desirable" = 11.5%) and the Flagger limit is 30% of the long-term laboratory median.

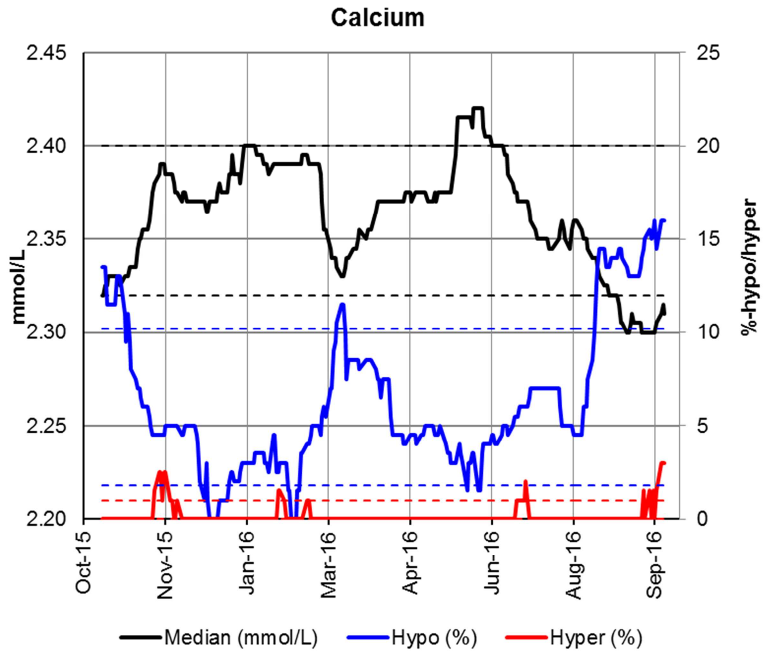


Figure 4. Similar to Figure 1, but for calcium. Note, the Percentiler limit is 0.04 mmol/L (= 1.7%; "desirable" = 0.8%) and the Flagger limit is 70% of the long-term laboratory median.

Sodium

Figure 5 shows a drift in sodium values from ~140 mmol/L up to ~141.4 mmol/L, followed by a downwards shift to ~139 mmol/L, and a normalization back to 141.5 mmol/L. The hypo flagging rate is affected by the shift and changes from ~2% to ~5%, and normalizes back to ~2%. The hyper flagging rate is affected by the drift as it increases from ~3% up to ~12%; due to the downward shift it drops from ~12% to nearly 2% and then increases back to ~12%. Operating the test within the ±1 mmol/L limit would result in hyper flagging rates between 2 and 12%.

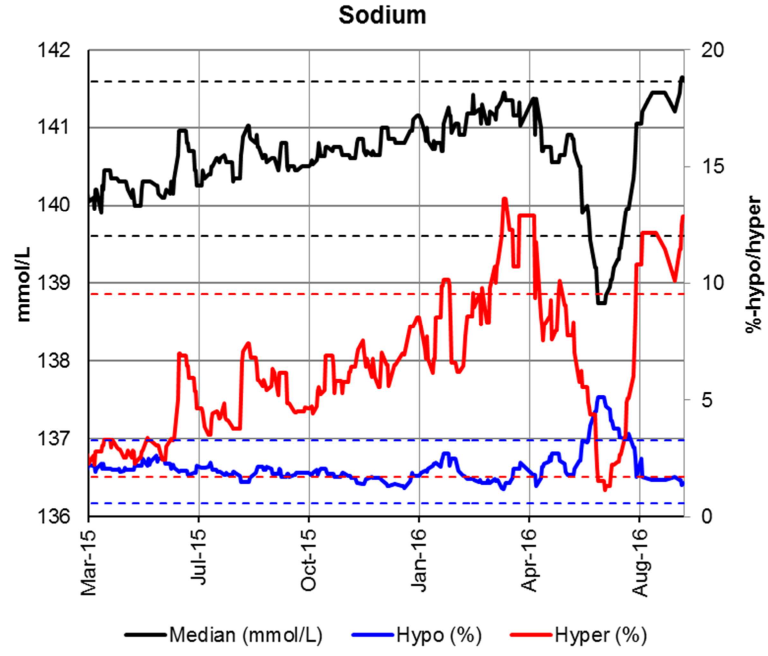


Figure 5. Similar to Figure 1, but for sodium. Note, the Percentiler limit is 1 mmol/L (= 0.7%; “desirable” = 0.2%) and the Flagger limit is 70% of the long-term laboratory median.

Discussion

The long-term experience of our Flagger and Percentiler applications allowed the investigation of the relationship between analytical performance specifications and surrogate medical decisions, i.e., flagging rates against locally used reference interval or decision limits. However, it is to note that the Flagger and Percentiler applications monitor laboratory data at different concentration levels. Therefore, analytical variation observed in the Percentiler may not directly reflect change in flagging rates. Nevertheless, most examples in this publication show that the most prominent flagging rate (be it hypo or hyper) correlates quite well with analytical variations seen in the Percentiler application.

We could demonstrate that restricting the increase in flagging rates to 20% is realistic for some of the most important analytes (total-cholesterol and glucose) despite the fact that their relatively low biological variation requires ambitious analytical stability limits (2.3% for glucose, for example; see Table 1). This contrasts with the phosphate test (analytical stability limit of 3.4%), where we needed to operate the Flagger with a 50% limit. We speculate that this is because manufacturers (and laboratories) give special attention to analytes that are in the public or scientific focus which, unfortunately, is not the case for the phosphate test. We could apply 30% Flagger limits for most of the other tests which nicely corresponds to analytical stability limits derived from biological variation and, consequently, limits applied in the Percentiler. For analytes with low biological variation (ALB, CL, PROT, CA, MG, NA) we had to set the Flagger limits higher (50% or 70%). Note, we also had to apply a 50% limit for uric acid because of its seasonal variation (somewhat higher in the summer). With the selected examples, we can give insight in the effect of analytical variation on locally used flagging rates.

The selected examples may demonstrate in more detail the utility of the Flagger/Percentiler combination. The total-cholesterol example (Figure 1) demonstrates the maturity of the test. Certain laboratories were able to keep the test within the 20% Flagger limit for over 2 years, corresponding to an analytical stability within ±0.2 mmol/L (desirable stability = 4.1%). Similar holds true for the glucose test which, even, requires better stability (desirable stability = 2.3%) because of its lower biological variation as compared to total-cholesterol (Figure 2). ALT tests, typically, were stable within the recommended limits (not shown), except the one of Ortho Clinical Diagnostics used to construct Figure 3 and demonstrating that the analytical instability gave significant changes in flagging rates (hypo from 0% to 28%; hyper from 0% to 8%). Analytical stability in the order of 2 IU for concentrations in the reference interval could greatly support the utility of the ALT test for newer applications, such as early detection of metabolic changes (“metabolic syndrome”) (18-20). In that connection, cut-offs for high-normal may be useful as most laboratories, only, flag very high results. As reported earlier, calcium flagging rates are very much influenced by analytical instability because of the low biological variation (12). The hypo flagging rate was ~15% at 2.30 mmo/L, whereas it was only ~3% at 2.42 mmol/L in the example shown in Figure 4. This means that operation of the calcium test with an instrument or laboratory bias of ~0.12 mmol/L could result in five-fold different hypo flagging rates. For sodium, we confirmed our earlier observations (16). In the example of Figure 5, the hyper flagging range nearly triplicated (3% to 9%)

due to a drift from 140 to 141 mmol/L. Achieving a 1 mmol/L stability could significantly improve the utilization of the sodium test by clinicians (21, 22). Indeed, several laboratories were able to reach such a stability over more than 2 years, supporting the applicability of the 1 mmol/L stability which, however, still can result in significant increase in flagging rates.

The Flagger application is of limited value for low-throughput laboratories because of high variability of the flagging rates that cannot be compensated for by choosing a higher n for calculating the moving median. On the other hand, also those laboratories profit from the application simply by being part of it and learning from the other participants via the reports we regularly send. Currently, the number of participants is too low to compare flagging rates across laboratories or peer groups.

Future plans are the comparison of locally used cut-off points for flagging. Either, they may reveal justified reasons why different cut-offs are used or give the opportunity for harmonization when no reason can be identified.

Conclusion

Overall, we showed that the combination of monitoring flagging rates together with daily patient medians is capable of translating the effect of analytical variation into its effect on surrogate medical decisions. Our results re-iterated the utility of the concept of setting analytical performance specification from biological variation. The Flagger/Percentiler application can form a bridge of the “medium-level” hierarchy for setting analytical specifications (biology) to the “top-level” (clinical situations). The advantage of the application is the direct visualization of analytical instability at the laboratory level using data readily available in the laboratory itself.

References

1. Belk WP, Sunderman FW. A survey of the accuracy of chemical analyses in clinical laboratories. Am J Clin Pathol 1947;17(11):853-61.
2. Tonks DB. A study of the accuracy and precision of clinical chemistry determinations in 170 Canadian laboratories. Clin Chem 1963;9(4):217-33.
3. Cotlove E, Harris EK, Williams GZ. Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. 3. Physiological and medical implications. Clin Chem 1970;16:1028–32.
4. Stöckl D, Baadenhuijsen H, Fraser CG, Libeer JC, Petersen PH, Ricós C. Desirable routine analytical goals for quantities assayed in serum. Discussion paper from the members of the external quality assessment (EQA) Working Group A on analytical goals in laboratory medicine. Eur J Clin Chem Clin Biochem 1995;33(3):157-69.

5. Kallner A, McQueen M, Heuck C. The Stockholm Consensus Conference on quality specifications in laboratory medicine, 25-26 April 1999. Scand J Clin Lab Invest 1999;59(7):475-6.
6. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, Petersen PH, Schimmel H, Sikaris K, Panteghini M. 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(6):833-5.
7. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. Clin Chem Lab Med 2015;53(6):829-32.
8. Klee GG. Tolerance limits for short-term analytical bias and analytical imprecision derived from clinical assay specificity. Clin Chem 1993;39:1514–8.
9. Stepman HC, Stöckl D, Twomey PJ, Thienpont LM. A fresh look at analytical performance specifications from biological variation. Clin Chim Acta 2013;421(6):191-2.
10. De Grande LA, Goossens K, Van Uytfanghe K, Halsall I, Yoshimura Noh J, Hens K, Thienpont LM. Using "big data" to describe the effect of seasonal variation in thyroid-stimulating hormone. Clin Chem Lab Med 2016 Jul 22. pii: /j/cclm.ahead-of-print/cclm-2016-0500/cclm-2016-0500.xml. doi: 10.1515/cclm-2016-0500. [Epub ahead of print].
11. De Grande LA, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru MM, Thienpont LM; IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta. 2016 Apr 27. pii: S0009-8981(16)30157-7. doi: 10.1016/j.cca. 2016.04.032. [Epub ahead of print].
12. De Grande L, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont L. The Empower Project – A new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53(8):1197-204.
13. Goossens K, Van Uytfanghe K, Twomey P, Thienpont L, and Participating laboratories. Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make “Big Data” work. Clin Chim Acta 2015;445:12-18.
14. Goossens K, Brinkmann T, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Clin Chem Lab Med 2015;53(10):e269-70.
15. Van Houcke SK, Stepman HC, Thienpont LM, et al. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.

16. Stepman HC, Stöckl D, Stove V, et al. Long-term stability of clinical laboratory data: sodium as benchmark. Clin Chem 2011;57:1616-7.

17. Westgard QC. Biological variation database, and quality specifications for imprecision, bias and total error (desirable and minimum). The 2014 update. http://www.westgard.com/biodatabase-2014-update.htm (accessed October 2016).

18. Liu Z, Que S, Xu J, Peng T. Alanine Aminotransferase-Old Biomarker and New Concept: A Review. Int J Med Sci 2014;11(9): 925-35.

19. Oh RC, Hustead TR. Causes and Evaluation of Mildly Elevated Liver Transaminase Levels. Am Fam Physician 2011;84(9):1003-8.

20. Aragon G, Younossi ZM. When and how to evaluate mildly elevated liver enzymes in apparently healthy patients. Cleve Clin J Med 2010;77(3):195-204.

21. Solinger AB, Rothman SI. Risks of mortality associated with common laboratory tests: a novel, simple and meaningful way to set decision limits from data available in the Electronic Medical Record. Clin Chem Lab Med 2013;51(9):1803-13.

22. Wald R, Jaber BL, Price LL, Upadhyay A, Madias NE. Impact of hospital-associated hyponatremia on selected outcomes. Arch Intern Med 2010;170(3):294-302.

Supplemental Figures 2-18

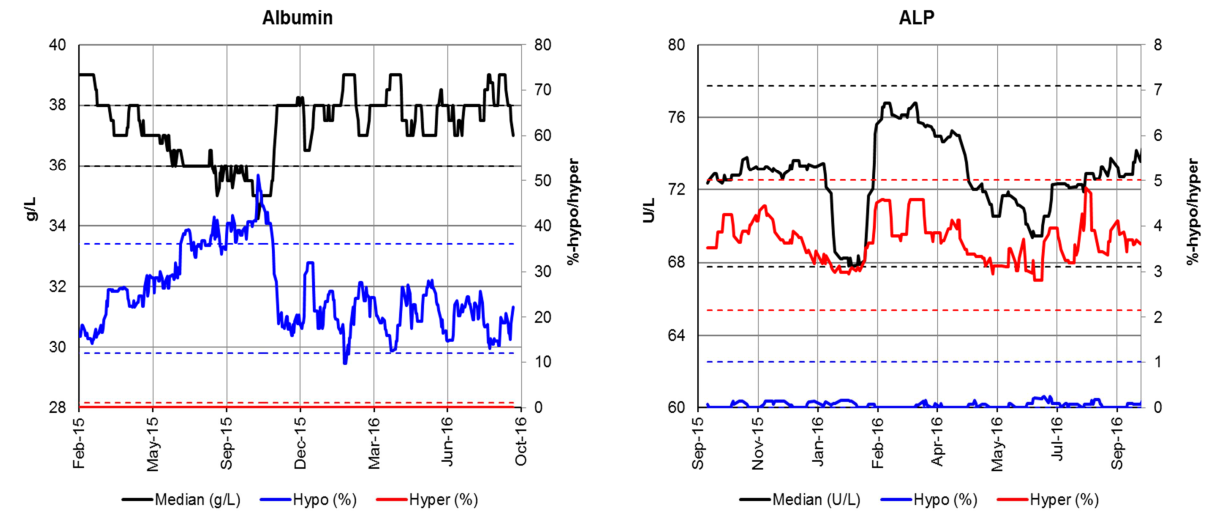


Figure 2S and 3S. Albumin and ALP.

Figure 2S. The figure shows a drift of albumin values from ~ 39 g/L to ~35 g/L, followed by a shift back to ~38 g/L. The hypo flagging rate increases from ~20% to ~42% and falls back to ~20%. This laboratory has, like most laboratories, no hyper flagging for albumin. Note, the Percentiler limit is 1 g/L (= 2.3%; “desirable” = 1.4%) and the Flagger limit is 50% of the long-term laboratory median.

Figure 3S. ALP assays are among the most stable assays across laboratories and manufacturers. The figure here shows several moderate shifts of the alkaline phosphatase values, first the values drop from ~73U/L to ~68 U/L, second they go up to ~76 U/L, third they fall to ~70 U/L and fourth return to ~73U/L. The hypo flagging rate is very low, as mostly is the case. The hyper flagging rate varies concordantly with the analytical shifts between ~3 % to ~5%. Note, the Percentiler limit is 5 U/L (= 6.8%; “desirable” = 6.7%) and the Flagger limit is 30% of the long-term laboratory median.

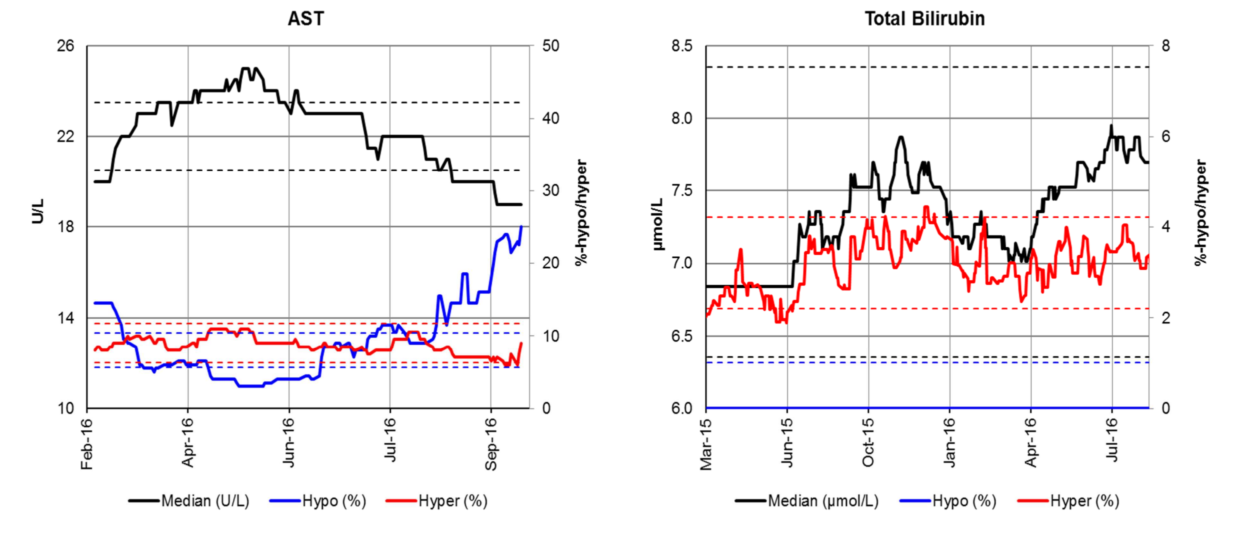


Figure 4S and 5S. AST and total-bilirubin.

Figure 4S. The figure shows AST values varying from ~20 U/L to ~24 U/L and back to ~20 U/L. The hypo flagging rate decreases concordantly with the analytical variation from ~13% to ~4% and increases again to ~22%. The hyper flagging rate slightly drops from ~8% to ~6%. Note, the Percentiler limit is 1.5 U/L (= 6.5%; “desirable” = 6.5%) and the Flagger limit is 30% of the long-term laboratory median.

Figure 5S. The total-bilirubin assay is among the most stable assays across manufacturers and laboratories; typically, the limits are never violated. The figure shows total bilirubin values shifting from ~6.8 µmol/L up to ~7.5 µmol/L; then they decrease to ~7.0 µmol/L followed by an increase to ~7.6 µmol/L. The hyper flagging rate is concordantly affected by both analytical shift as it goes from ~2.5% up to nearly 4% and comes back to ~3%. As expected, the hypo flagging rate is almost zero %. Note, the Percentiler limit is 1 µmol/L (= 10%; “desirable” = 9.0%) and the Flagger limit is 30% of the long-term laboratory median.

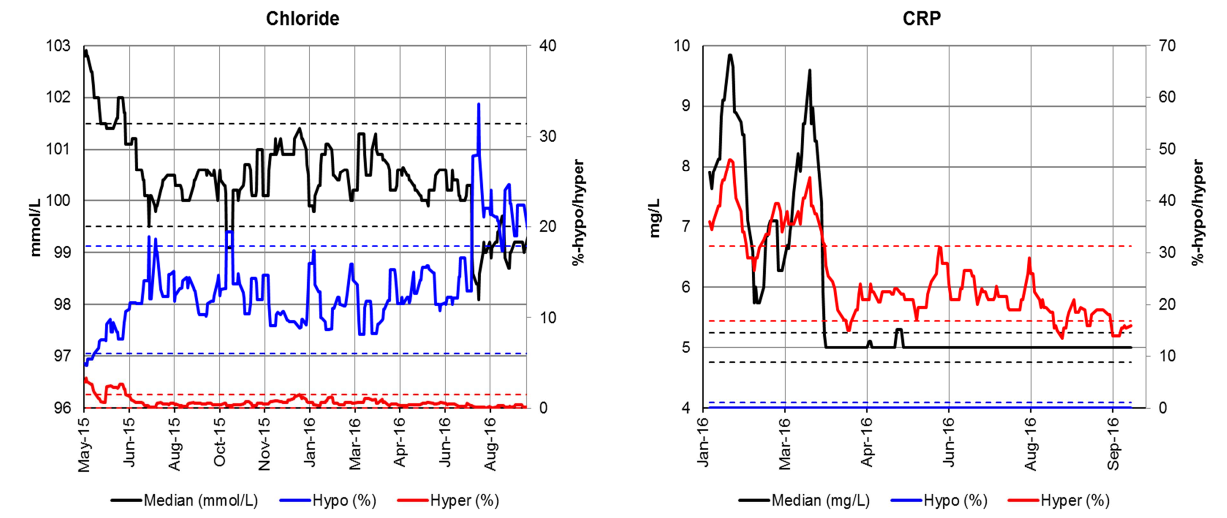


Figure 6S and 7S. Chloride and CRP.

Figure 6S. The figure shows two shifts in the chloride values, i.e., from ~ 103 mmol/L to ~100.5 mmol/L, and to ~99 mmol/L. The hypo flagging rate is most influenced by both analytical shifts as it increases from ~5% to ~21%. The hyper flagging rate decreases by the first downward shift from ~2% to nearly 0%. Note, the Percentiler limit is 1 mmol/L (= 1.0%; “desirable” = 0.5%) and the Flagger limit is 50% of the long-term laboratory median.

Figure 7S. The figure shows four major shifts in the CRP values, i.e., from ~7.0 mg/L to ~ 9.9 mg/L, then down to ~5.8 mg/L, up to ~9.5 mg/Land down to ~5 mg/L. The hyper flagging rate is influenced by the analytical shifts as it first increases from ~35% to ~48%, then drops to ~28% to increase again to ~35-45%, and then decreases to ~15-20%. The hypo flagging rate is and remains zero as expected. Note, the Percentiler limit is 0.25 mg/L (= 9.6%; “desirable” = 21.8%) and the Flagger limit is 30% of the long-term laboratory median.

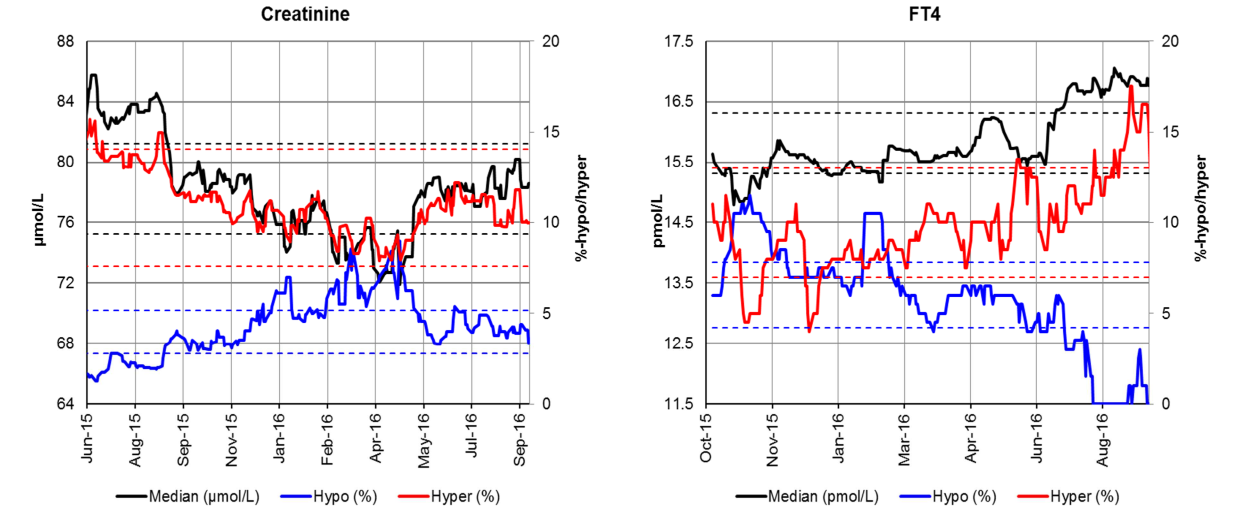


Figure 8S and 9S. Creatinine and FT4.

Figure 8S. The figure shows drifting creatinine values from ~84 µmol/L down to ~74 µmol/L, followed by a shift to~78 µmol/L. The hypo flagging increases concordantly with the drift from ~3% to ~7.5% to normalize again to ~3%. The hyper flagging rate gradually decreases from ~15% down to nearly 8% and then normalizes to ~11%. Note, the Percentiler limit is 3 µmol/L (= 3.9%; "desirable" = 4.0%) and the Flagger limit is 30% of the long-term laboratory median.

Figure 9S. The figure shows FT4 values drifting and shifting from 15.5 pmol/L up to 17 pmol/L. The hypo flagging rates decrease from ~10% to ~0%, while the opposite happens for the hyper flagging rate (increases from ~8% to nearly 16%). Note, the Percentiler limit is 0.5 pmol/L (= 3.3%; "desirable" = 3.3%) and the Flagger limit is 30% of the long-term laboratory median.

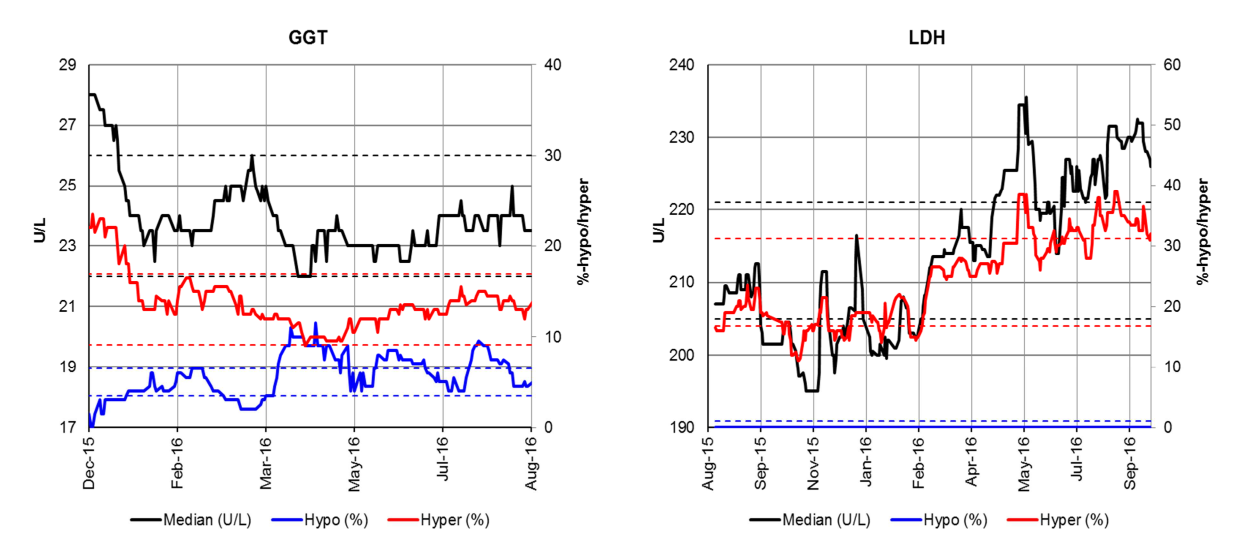


Figure 10S and 11S. GGT and LDH.

Figure 10S. The figure shows a downward shift in GGT values from ~28 U/L to ~24U/L. The hypo flagging rate increases concordantly with the analytical shift from ~ 3% to ~10%. The hyper flagging rate drops from ~22% to nearly 10%. Note, the Percentiler limit is 2 U/L (= 9.1%; “desirable” = 11.1%) and the Flagger limit is 30% of the long-term laboratory median.

Figure 11S. The figure shows drifting LDH values from ~200 U/L up to to ~230 U/L. The hyper flagging rate gradually increases from ~18% to nearly 33%. As is the case for other enzymes, the hypo flagging rate is nearly 0%. Note, the Percentiler limit is 8 U/L (= 4.6%; “desirable” = 4.3%) and the Flagger limit is 30% of the long-term laboratory median.

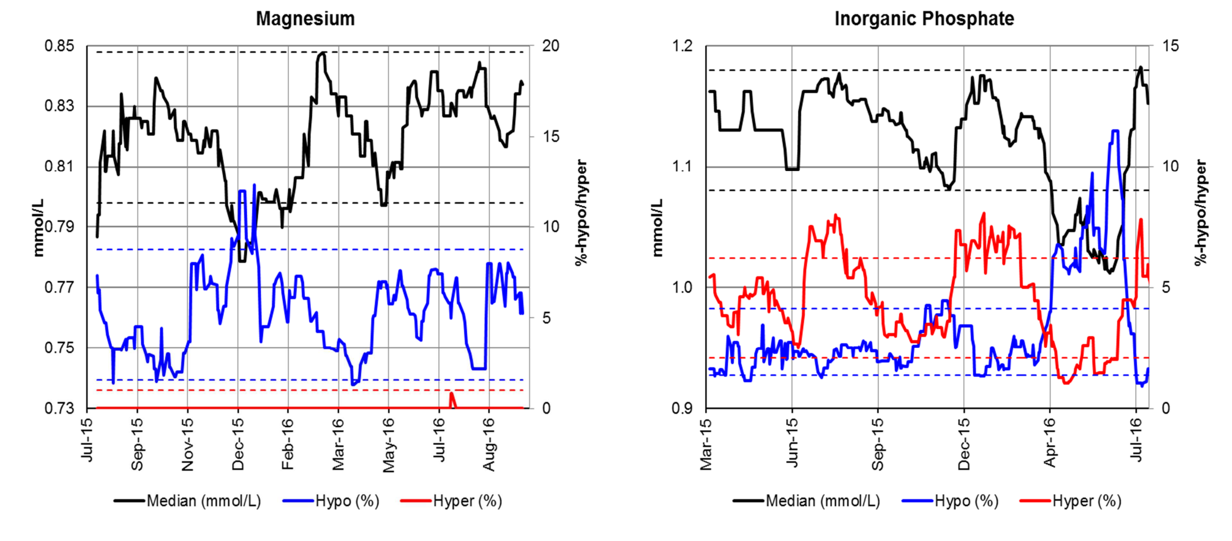


Figure 12S and 13S. Magnesium and inorganic phosphate.

Figure 12S. The figure shows several shifts in the magnesium values, i.e., a first drop from ~0.83 mmol/L to ~0.79 mmol/L, followed by an increase to ~0.84 mmol/L, to drop back to ~0.81 mmol/L and finally increase to 0.835 mmol/L. The hypo flagging rate decreases from ~7.5% to ~2.5%, then returns to ~7.5-9.0%, comes down to ~2.5%, to normalize again to ~7.5%. The hyper flagging rate is nearly 0%. Note, the Percentiler limit is 0.02 mmol/L (= 3.0%; “desirable” = 1.8%) and the Flagger limit is 70% of the long-term laboratory median.

Figure 13S. The figure shows several shifts in the inorganic phosphor values, the major ones being from ~1.13 mmol/L to ~1.03 mmol/L and back to ~1.17 mmol/L. The hypo flagging rate concordantly increases from ~2% to ~12% and drops back to ~2%. The hyper flagging rate mainly decreases due to the downwards analytical shift from ~7% to nearly 2.5%. Note, the Percentiler limit is 0.05 mmol/L (= 4.4%; “desirable” = 3.4%) and the Flagger limit is 50% of the long-term laboratory median.

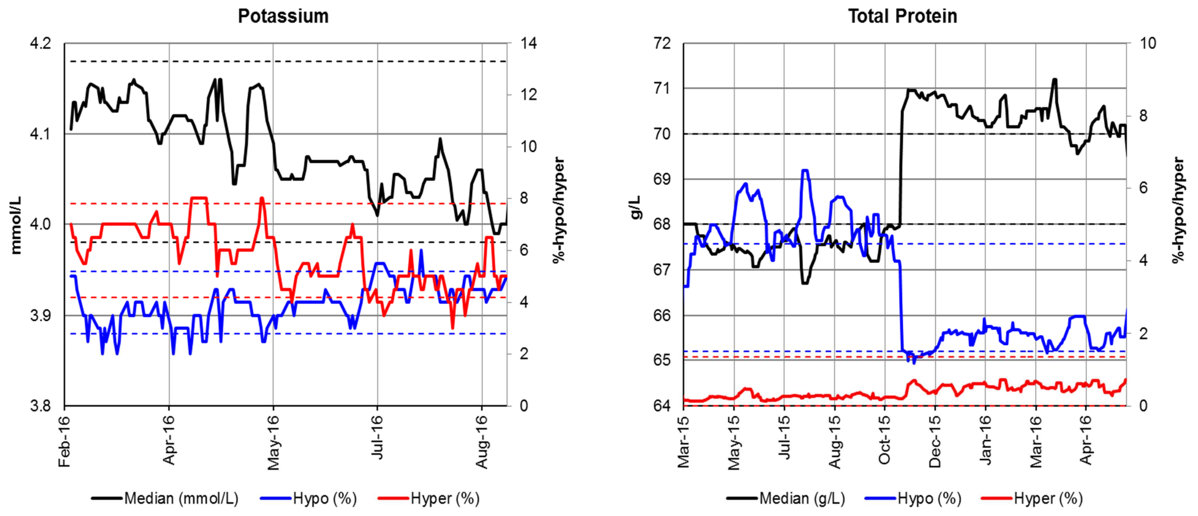


Figure 14S and 15S. Potassium and total-protein.

Figure 14S. The potassium assay is among the most stable assays across manufacturers and laboratories; typically, the limits are never violated. The figure shows potassium values going down from ~ 4.15 mmol/L to ~4.05 mmol/L. The hypo flagging rate increase concordantly from ~3% to ~5%, while the inverse happens for the hyper flagging rate decreasing from ~7% to nearly 4%. Note, the Percentiler limit is 0.1 mmol/L (= 2.4%; “desirable” = 1.8%) and the Flagger limit is 30% of the long-term laboratory median.

Figure 15S. The figure shows a major shift in the total protein values from ~67.5 g/L to ~70.5 g/L. The hypo flagging rate decreases concordantly from ~5% to ~2%. The hyper flagging rate triplicates (from ~0.2% to nearly 0.6%), however, is generally low. Note, the Percentiler limit is 1 g/L (= 1.4%; “desirable” = 1.4%) and the Flagger limit is 50% of the long-term laboratory median.

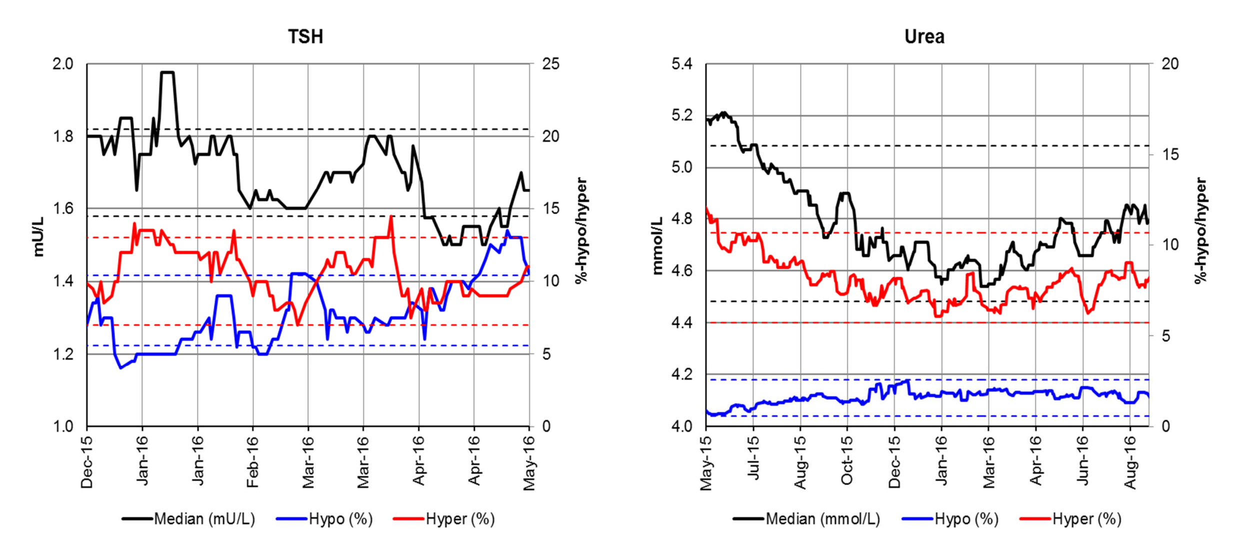


Figure 16S and 17S. TSH and urea.

Figure 16S. The figure shows variable TSH values ~1.8 mIU/L to ~1.55 mIU/L. The hypo flagging rate follows the trend grossly and varies between ~5% and 12%. The hyer flagging rate follows the analytical trend nicely and varies between ~7% and 13%. Note, the Percentiler limit is 0.12 mmol/L (= 7.7%; “desirable” = 7.8%) and the Flagger limit is 30% of the long-term laboratory median.

Figure 17S. The figure shows drifting urea values from ~5.2 mmol/L to ~4.6 mmol/L and, finally slightly increasing to ~4.8 mmol/L. The hypo flagging rate <1% gradually increases up to ~2.5% and levels off at ~1.5%, while the hyper flagging rate decreases from ~12% to nearly 7.5%. Note, the Percentiler limit is 0.3 mmol/L (= 6.0%; “desirable” = 5.6%) and the Flagger limit is 30% of the long-term laboratory median.

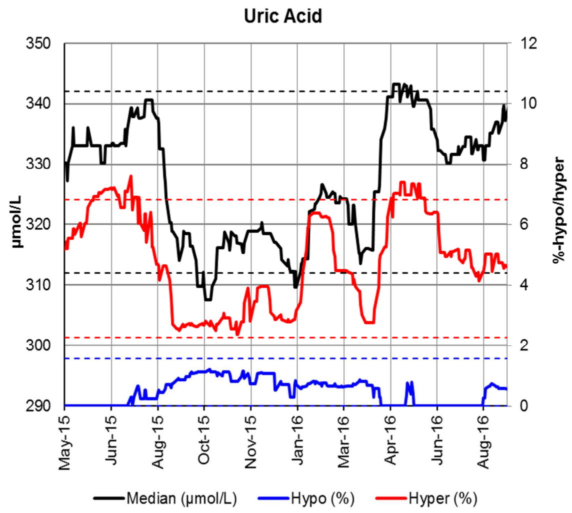


Figure 18S. Uric acid.

Figure 18S. The figure shows a downward and upward shift in the uric acid values from ~335 µmol/L to ~315 µmol/L and back to ~340 µmol/L. Due to the downward analytical shifts, the hyper flagging rate at ~7% falls to ~2.5%; it returns to nearly 7% after the upwards shift. As expected, the hypo flagging rate is very low and little affected by the variation and varies between 0 and 1%. Note, the Percentiler limit is 15 µmol/L (= 4.8%; "desirable" = 4.9%) and the Flagger limit is 50% of the long-term laboratory median.

4 Chapter IV

Master Comparisons

This chapter is based upon:

- Goossens K, Stöckl D, Van Uytfanghe K, Thienpont L. Hierarchy of target values and acceptance limits in EQA. IFCC Worldlab – 22nd international congress of clinical chemistry and laboratory medicine (22-26 June 2014), Istanbul, Turkey & EuroLabFocus – 3rd EFLM-UEMS congress (7-10 October 2014), Liverpool, UK [poster].
- Goossens K, Stöckl D, Van Uytfanghe K, Thienpont L. Master Comparisons 2014: Frozen, single donation sera for External Quality Assessment of in-vitro diagnostic assays and clinical laboratories - General Report: Assay and Laboratory Performance. 18 August 2014 [report].
- Goossens K, Stöckl D, Van Uytfanghe K, Thienpont L. Master Comparisons 2014: Frozen, single donation sera for External Quality Assessment of in-vitro diagnostic assays and clinical laboratories - Laboratory Guide for Data Interpretation. 18 August 2014 [report].
- Goossens K, Thienpont L. Reference measurement procedure corrected all method trimmed mean - The best of two worlds. Clin Chim Acta 2015;440:55–56.
- Goossens K, Van Uytfanghe K, Thienpont L. Trueness and comparability assessment of widely used assays for 5 common enzymes and 3 electrolytes. Clin Chim Acta 2015;442:44-5.

Highlights

Comparability of measurement results can best be achieved by establishment of traceability to a higher order method. If those methods are absent, harmonization to a lower-order reference material or to a surrogate target, such as the All Manufacturer Trimmed Mean (AMTM), is the best alternative. To maintain comparability between routine methods there is, amongst others, a constant need for QC in the clinical laboratory. EQA has proven to be a valuable tool to address that QC, however, it remains limited to the objectives that can be reached because of the use of non-commutable materials. This is why we developed our Master Comparison initiative in order to assess the comparability status of five commonly measured enzymes and three electrolytes by use of commutable, single donation serum samples.

Highlights of the research:

- Reference Measurement Procedures (RMP's) are used to demonstrate the standardization status of eight routine procedures for ALT, AST, LDH and GGT.
- Reference Measurement Procedure - corrected All Manufacturer Trimmed Mean (RMP-corrected AMTM) target values are superior over RMP values when investigating causes of dispersion of results from routine procedures.
- Peer performance and assay comparability were generally good for chloride, potassium and sodium. An exception was the negative bias of the Roche assays for chloride.
- For enzyme assays both manufacturers and laboratories still have a lot to do to accomplish traceability of measurement results.
    - There is a need to phase out the non-IFCC assay variants.
    - The quality, comparability, and standardization of the ALT assays needs to be improved.
    - GGT and LDH assays require better comparability.
    - There exists a general need for improving peer comparability at low concentration levels.
    - Nearly all enzymes studied require single-manufacturer efforts for improvement of comparability.

1. Introduction

One of the main concerns in laboratory medicine is a lack in comparability between different assays, different laboratories, and even between different instruments in a single laboratory. If this problem occurs patient data cannot be exchanged between different sites or even between different instruments on a single site. Non-comparability is not only caused by the use of different methods or assays and different instrumentation, it can also find its origin in the use of different calibration material and the occurrence of commutability issues (1). In order to avoid non-comparability, the establishment of traceability to a higher order method has been proposed. However, in practice, this is not always achievable, simply because of the non-existence of higher order methods/material. In other cases, the implementation of existing traceability chains can fail in achieving the comparability aimed at (2).

A major tool to address the current status of comparability between different assays and laboratories is the compulsory participation of laboratories to EQA surveys. These mostly make use of proficiency testing materials which are measured by each participant in the same manner as done for their patient samples (3-6). If these materials are commutable with the examined assays, they can be used to compare inter-assay results. However, if they are not, which is quite often the case, then a false impression about the trueness of laboratory results might be introduced (7-8). Therefore, non-commutable materials should only be used to address comparability between instruments/laboratories within the same peer group.

In our Master Comparison studies we make use of single donation serum samples from 20 apparently healthy donors. These samples are as commutable as possible and are therefore ideally suited to address comparability between both the laboratories within a single peer group and across assays (9). In addition, by measuring several analytes with a RMP we were also able to demonstrate the standardization status of the different assays. However, our approach has some limitations: because we restricted our method design to apparently healthy donors, our results don't show the comparability/standardization status over the entire concentration range (i.e. including “diseased” samples). In addition, because the serum volume available from a single donor is limited, our approach to EQA is only suited to serve a small number of laboratories.

Previously performed Master Comparison studies assessed the comparability status of albumin, calcium, creatinine, glucose, HDL-cholesterol, LDL-cholesterol, magnesium, phosphate, total protein, total cholesterol, triglycerides, and uric acid (10-12). In our 2014 study we focused our efforts on five common enzymes and three electrolytes.

2. Study design

The Master Comparisons are EQA surveys performed with 20 freshly frozen, single donation sera from apparently healthy volontueers. In the 2014 survey eight analytes were assessed in eight different peer groups. The studied analytes were: ALP, ALT, AST, LDH, GGT, potassium, sodium, and chloride. Participants were selected to obtain carefully controlled peer groups for Abbott Architect (n = 21), Beckman AU (n = 19), Beckman DxC (n = 11), Ortho Vitros (n = 19), Roche Cobas (n = 26), Roche Modular (n = 9), Siemens Advia (n = 12), Siemens Vista (n = 8), and Siemens Xpand (n = 1) systems. Also manufacturers participated in the study with a minimum of three systems in their central application laboratories. Because different variants were used for the measurement of certain enzymes, it was not always possible to establish a peer group. For example, ALT and AST assays are performed either with or without pyridoxale phosphate (PP) activation, and for the measurement of LDH there are two method principles available on the market. One method applies the forward reaction (lactate to pyruvate), the other the reverse reaction (pyruvate to lactate).

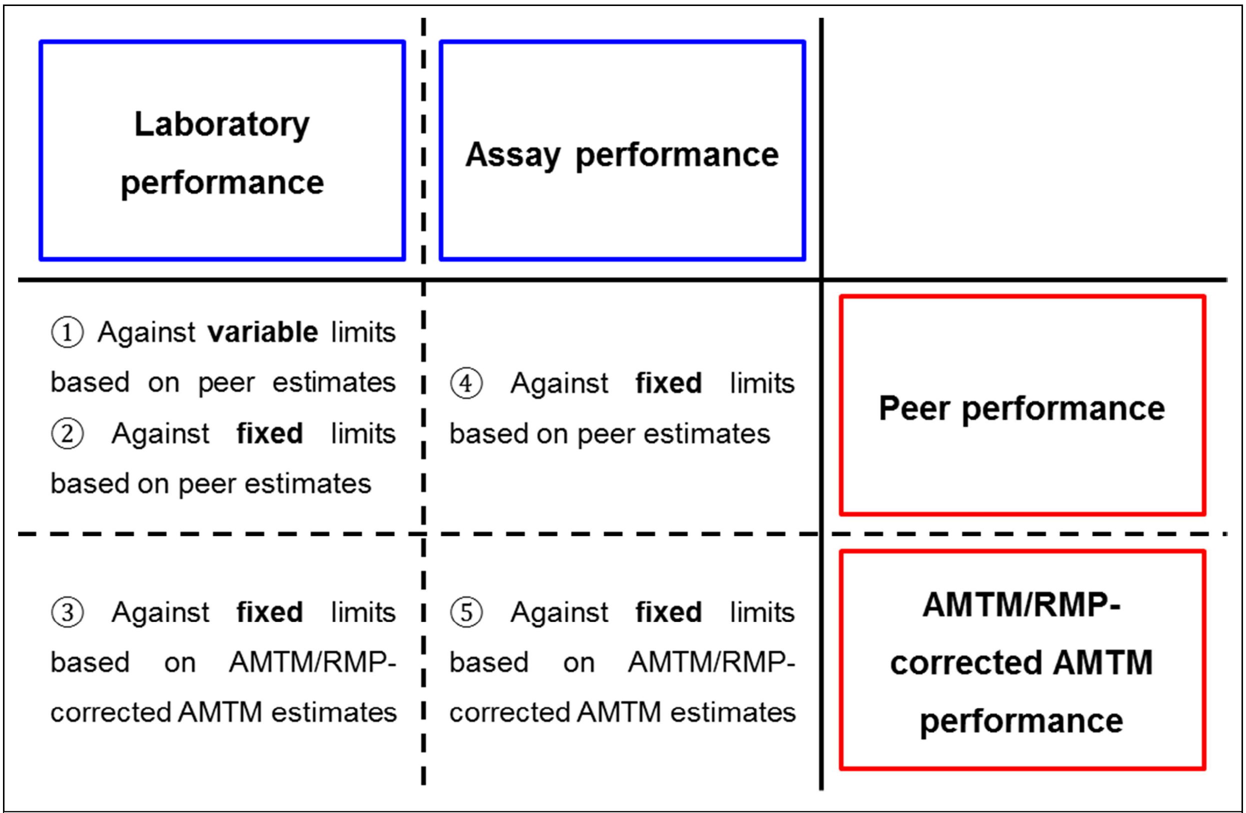
Study objectives

The objectives of the study were to assess both assay and laboratory performance, as well as their comparability at the peer group level and at the “reference level”. The latter is obtained by either comparing the measurements of the laboratories to the Reference Measurement Procedure -corrected All Manufacturer Trimmed Mean (RMP-corrected AMTM) or to the AMTM (the concept behind these targets is explained on page 150).

Assessment criteria and performance limits

Assessment criteria or quality indicators (QI) were i) %-Sy/x from linear regression as measure for within-run imprecision (laboratory versus the peer group) and as measure for combined imprecision and sample-related effects (laboratory versus the AMTM/RMP-corrected AMTM, note: this measure is also affected by the AMTM/RMP-corrected AMTM uncertainty); ii) bias versus the peer and AMTM/RMP-corrected AMTM at three concentration levels; iii) peer and AMTM/RMP-corrected AMTM TE; iv) peer group variation at three concentration levels (peer group coefficient of variation, CV).

In total, the laboratory quality is judged against two targets with three types of limits, and the manufacturer quality against two targets with two types of limits (shown schematically in Figure 1). Performance limits were hierarchically structured as laboratory performance versus (1) variable limits based on peer group estimates; (2) fixed limits based on peer group estimates; (3) fixed limits based on AMTM/RMP-corrected AMTM estimates; and assay performance versus (4) fixed limits based on peer group estimates; and (5) fixed limits based on AMTM/RMP-corrected AMTM estimates.



Laboratory performance

Assay performance

① Against **variable** limits based on peer estimates

② Against **fixed** limits based on peer estimates

④ Against **fixed** limits based on peer estimates

Peer performance

③ Against **fixed** limits based on AMTM/RMP-corrected AMTM estimates

⑤ Against **fixed** limits based on AMTM/RMP-corrected AMTM estimates

AMTM/RMP-corrected AMTM performance

Figure 1. Hierarchy of target values and acceptance limits. The laboratory quality is judged against two targets with three types of limits (①, ②, and ③), and the manufacturer/assay quality against two targets with two types of limits (④, and ⑤).

3. Hierarchy of target values and decision limits

Targets

A) Peer target

Peer group target values were calculated as the mean from laboratories, except the outlying laboratories. The latter were identified by a two-sided Grubbs test (13), based on 95% probability (20 out of 975 tests = 2.1%). These outliers gave no indication for laboratory or assay quality, but were mainly related to peer groups with small variation (for example, four labs for chloride and three labs for sodium).

B) AMTM target

AMTM target values were calculated as the mean of the peer group targets, except the outliers. The latter were also identified by a Grubbs test, but no outliers were detected. However, for chloride, the Cobas and Modular were not included in the AMTM because the suspected difference versus the other assays was supported by measurements with a RMP.

C) RMP-corrected AMTM target

The 20 serum samples were measured with RMPs for ALT, AST, GGT, and LDH (14). Target values were assigned by the WEQAS Reference Laboratory in Cardiff (15). The laboratory is accredited and participates in the IFCC ring trials for reference laboratories (16). Measurements were made in triplicate over four days. Median CV of the RMPs were 4.8% (ALT), 2.9% (AST), 3.3% (GGT), and 1.8% (LDH). Supplementary RMP values for chloride (10 additional samples) were determined in the INSTAND (Düsseldorf) reference laboratory (17). The laboratory is accredited and listed in the Joint Committee for Traceability in Laboratory medicine (JCTLM) database (18).

The RMP values were used to correct the AMTM target, to include the bias information from the trueness-based methods. This was done to avoid potentially high dispersion of the data in method comparison analysis with the RMP as a target. RMPs often apply instrumental analytical procedures which require extensive sample clean-up, and involve manual steps. This makes them vulnerable to increased imprecision, low throughput and creates high measurement cost. This leads to measuring a low number of replicates. Routine procedures, on the other hand, are characterized by very low within-run measurement imprecision and are relatively cheap, which favors replication. Method comparison studies with RMPs with several routine procedures have the potential of combining “the best of two worlds”, namely, high accuracy provided by the RMPs and low dispersion of the individual results provided by the AMTM.

This is demonstrated in Figure 1 of Article 4 for the GGT case, which shows the %-differences of two platforms versus the RMP and versus the RMP-corrected AMTM. Data distribution and R^2-values indicate the superiority of the RMP-corrected AMTM values over the RMP-values themselves. The dispersion of the data in the RMP comparison can be mainly attributed to the uncertainty of the RMP-values (median CV 3.3%). In consequence, potential random sample-related effects cannot be uncovered due to the overriding imprecision component of the RMP. The comparison with the RMP-corrected AMTM would suggest the absence of sample-related effects in the current examples. Note, while the bias information is not relevant for the assessment of random sample-related effects, it is conserved in the RMP-corrected AMTM because of the high quality of the regression equation between the AMTM and the RMP procedure (r = 0.995). An overview of these regression equations can be found in the annex (Supplemental Figure 1; pages 184-185).

In conclusion, the RMP-corrected AMTM target gives the same bias information as the RMP one, but has the advantage of lower distribution. We, therefore, chose the RMP-corrected AMTM as target value. Note, a more extensive explanation on the RMP-corrected AMTM can be found in the article on pages 168-170.

Decision limits for assay performance

A) Peer group estimates and limits

Peer group within-run imprecision was calculated as median laboratory peer Sy/x. Peer group variation (calculated as CV) was estimated at the concentration range ends (low and high), and at the mean concentration. Assay peer performance was assessed versus fixed limits which reflect the state-of-the-art performance of the assays (Table 1). Note, testing (estimates versus limits) was done without confidence intervals. The reasoning for choosing the respective values is outlined below.

Table 1 Assay fixed limits for peer group estimates

	ALP	ALT*	AST*	GGT	LDH*	Cl	K	Na
Sy/x	3.0	5.0	5.0	3.0	3.0	1.0	1.0	1.0
CV (biology)	3.2	9.7	6.2	6.7	4.3	0.6	2.3	0.3
Peer CV	5.0	7.5	5.0	7.5	5.0	1.5	2.0	1.0
Peer TE	11.4	17.7	13.9	15.8	11.4	3.5	4.4	2.8

* Limits are used for both ALT/AST with and without PP, and for both LDH forward and reverse

1. *Peer Sy/x*

Peer Sy/x (%) is a measure for within-run imprecision. It is calculated by performing ordinary lineair regression of the 20 samples for each laboratory against its peer target. The median of those peer Sy/x values was compared against the limits. Chloride, potassium, and sodium assays typically have within-run CVs of 0.5%, therefore the limits were chosen at 1% for all three. We aimed at 3% for the enzymes, however, we had to expand the values to 5% for ALT and AST. The table also contains the desirable total CV derived from desirable biological variation (19).

2. *Peer CV*

Due to the low biological variation, chloride and sodium (potassium to a lesser extent) require a tight control of bias in the laboratory. This is reflected in the generally good laboratory comparability for these tests and allows limits in the order of 1-2%. Therefore, limits of 1.5% for chloride, 2% for potassium, and 1% for sodium were chosen. According to the biological variation, the enzymes would tolerate higher variation, however, state-of-the-art comparability should be in the order of 5%. This could be applied for ALP, AST, and LDH. The limits for ALT and GGT had to be expanded to 7.5% because of the low concentrations (ALT) and/or possible lot variation.

3. *Peer TE*

The peer group TE was calculated as $1.96\sqrt{\mathrm{CVpeer}^2 + \mathrm{median\ Sy/x}^2}$ because both QIs (peer CV and peer Sy/x) are variance components.

B) AMTM/RMP-corrected AMTM estimates and limits

Peer group AMTM/RMP-corrected AMTM Sy/x was calculated as median laboratory AMTM/RMP-corrected AMTM Sy/x. The latter is calculated by performing ordinary lineair regression of the 20 samples for each laboratory against its AMTM/RMP-corrected AMTM target. The peer group AMTM/RMP-corrected AMTM bias was calculated as %-difference of a peer group from the AMTM/RMP-corrected AMTM target at three concentration levels; the respective confidence intervals were calculated from the respective laboratory biases. Peer group AMTM/RMP-corrected AMTM TE was calculated as: TE = Bias + 1.645 * Sy/x. Different from above, bias is not treated as variance component; the z-multiplier for 95% probability is chosen one-sided because of the bias component.

AMTM/RMP-corrected AMTM limits (Table 2) are based on the peer limits. The Sy/x limits are the same, while the bias limits are calculated from the peer CV limits plus the AMTM/RMP-corrected AMTM uncertainty; TE is calculated as Bias + 1.645.Sy/x. The table also contains desirable bias derived from biological variation (19) expanded by the AMTM/RMP-corrected AMTM uncertainty.

Table 2 Assay fixed limits for AMTM/RMP-corrected AMTM estimates

	ALP	ALT*	AST*	GGT	LDH*	Cl	K	Na
Sy/x	3.0	5.0	5.0	3.0	3.0	1.0	1.0	1.0
Bias	10.2	9.5	7.0	9.5	7.0	1.9	3.0	1.4
Bias (biology)	11.9	13.5	8.5	13.1	6.3	0.9	2.8	0.6
Peer TE	15.1	17.7	15.2	14.4	11.9	3.5	4.6	3.0
AMTM/RMP-corrected AMTM unc.	5.2	$2^{\#}$	$2^{\#}$	$2^{\#}$	$2^{\#}$	0.4	1	0.4

* Limits are used for both ALT/AST with and without PP, but only for LDH forward

$^{\#}$ Estimates refer to RMP-corrected AMTM targets

Decision limits for laboratory performance

A) Peer group estimates and variable limits

Laboratory within-run imprecision was calculated as %-Sy/x obtained from ordinary linear regression (laboratory versus peer group). Laboratory peer bias was calculated at three concentration levels (low, mean, high) as %-deviation from the peer group mean. Laboratory peer TE was calculated as TE = Bias + 1.645 * Sy/x.

Peer group dependent laboratory limits were calculated as 2 * median peer Sy/x, 2 * median peer CV, and 1.645 * median peer TE (see Annex, Supplemental Table 1). Testing of the estimates versus the limits was done in “absolute” terms (estimate > limit).

B) Peer group targets and fixed limits

Assessment of peer performance was also done versus fixed limits based on the “state-of-the-art” performance of modern multichannel instruments (Table 3). For imprecision (Sy/x) the same limits were chosen as for testing assay performance. The limits for bias and TE were expanded by the peer group uncertainty for each analyte individually (indicated in the bottom row of Table 3). Testing of the estimates versus the limits was done in “absolute” terms (estimate > limit), except testing of laboratory peer bias, which was done by use of the confidence intervals of the regression line.

Table 3 Laboratory fixed limits for peer group estimates

	ALP	ALT*	AST*	GGT	LDH*	Cl	K	Na
Sy/x	3.0	5.0	5.0	3.0	3.0	1.0	1.0	1.0
Bias	6.4	13.5	7.2	9.8	7.1	2.0	2.5	1.3
TE	12.8	23.7	16.1	18.1	13.5	4.0	4.9	3.1
Peer uncertainty	1.4	6.0	2.2	2.3	2.1	0.5	0.5	0.3

* Limits are used for both ALT/AST with and without PP, and for both LDH forward and reverse

C) AMTM/RMP-corrected AMTM estimates and limits

Assessment of laboratory AMTM/RMP-corrected AMTM performance was done versus the same fixed limits as for the assays (see Table 2 above). Testing of the estimates versus the limits was done in “absolute” terms (estimate > limit), except testing of laboratory AMTM/RMP-corrected AMTM bias, which was done by use of the confidence intervals of the regression line (similar as for the peer group targets and fixed limits).

4. Main study outcomes – 2014 survey

Participants and instruments

A total number of 126 participants contributed to the 2014 survey. Participants were recruited worldwide (Figure 2), but mostly Belgian laboratories participated (n = 69).

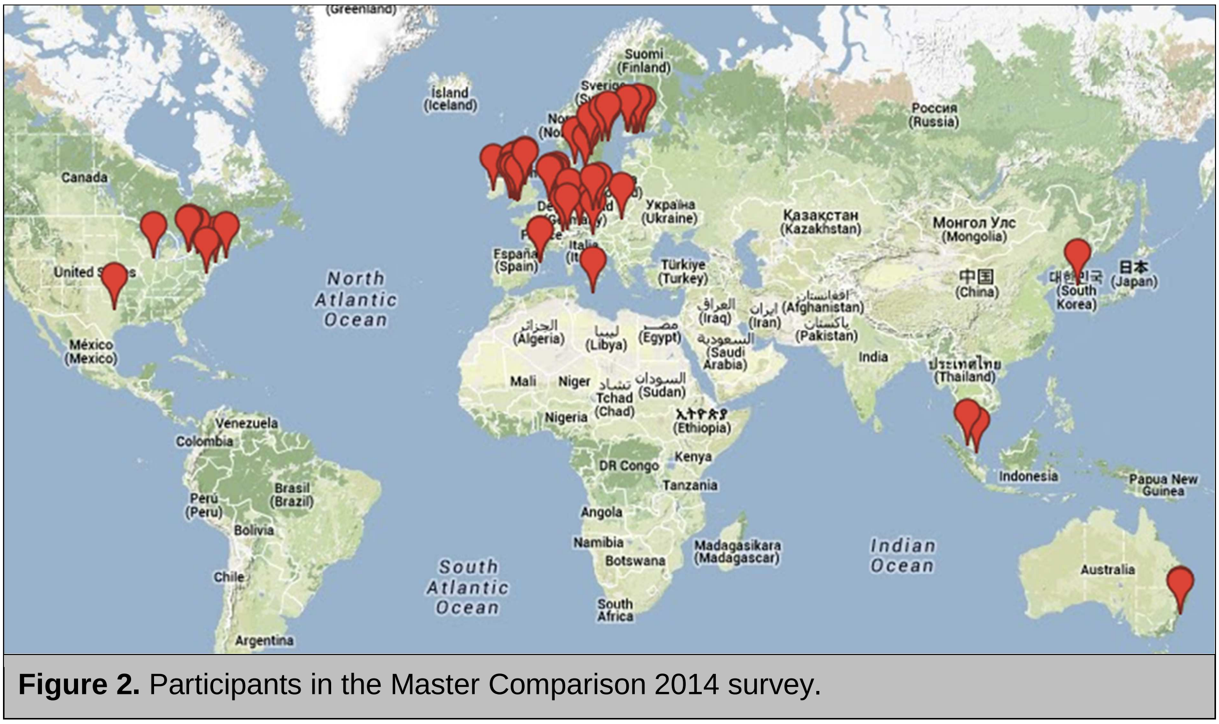


Figure 2. Participants in the Master Comparison 2014 survey.

With regard to the methods typically used in the participating laboratories, we made the following observations: for ALP nearly all laboratories apply the IFCC recommended method, however, a few “loners” still apply the Deutsche Gesellschaft für Klinische Chemie (DGKC) method (distributed over different manufacturers). For the latter no peer groups could be established. The IFCC recommended methods for ALT and AST are performed with PP activation. Many manufacturers, however, sell methods with and without PP activation. As standardization of enzyme measurements is based on establishing traceability to the IFCC recommended methods, all results are compared with the latter, also those that work without PP activation. In contrast, since the two method principles for the measurement of LDH, the forward and reverse reaction, give grossly different results, we treated LDH as two distinct analytes: LDH 1 (forward) and LDH 2 (reverse). LDH 1 results are compared to the RMP values while LDH 2 results are compared relatively to each other (note: the AMTM is not reliable due to its high uncertainty). Nevertheless, peer group estimates were available for LDH 2.

Table 4. Number of laboratories per specific assay (manufacturer/assay variant)

	Architect	AU	DxC	Cobas	Modular	Advia	Vista(E)	Vitros
ALP	21	16	10	26	9	9(D)	8	19
ALP DGKC	0	3	1	0	0	2	0	0
ALT + PP	2	9	2	5(A)	4	3(D)	8	19
ALT - PP	19	10	9	21	5	8	0	0
AST + PP	2	9	2	5(A)	4	3(D)	8	19
AST - PP	19	10	9	21	5	8	0	0
GGT	21	19	11	26	9	11(D)	8	19
LDH forward	21	9(B)	5	16	4(B)	5	8	0
LDH reverse	0	5(B)	6	10	4(B)	6(D)	0	19
Chloride	21	13(B)	11	25(C)	9	12	7(F)	19
Potassium	21	19	11	26	9	12	7(F)	19
Sodium	21	19	11	26	9	12	7(F)	19

Shaded fields: No peer group estimates because of low *n* and/or high variability.

(A) Backed-up by manufacturer data

(B) Some Beckman AU and Roche Modular laboratories did not report LDH and chloride

(C) One Roche Cobas used a radiometer for chloride

(D) One Siemens Advia used Beckman calibrators for the enzymes

(E) One Siemens participant used an Dimension Xpand

(F) One Siemens Vista laboratory did not report chloride, potassium and sodium

Typical peer group sizes were $n \geqslant 7$, however, we decided to calculate peer group estimates for $n < 7$ when i) results were backed up by manufacturer data, or ii) results fitted in the general expected picture of other assays (for example, LDH forward and reverse values).

Because of too few results, peer groups could not be established for ALP DGKC (all), ALT with PP (Architect, DxC, Modular, Advia), ALT without PP (Modular, Vista, Vitros), AST with PP (Architect, DxC, Modular, Advia), AST without PP (Modular, Vista, Vitros), LDH forward (Modular, Vitros), and LDH reverse (Architect, Vista). Note, for LDH reverse from Modular the peer group from Cobas was used.

Assay performance

A) Peer group performance

Peer group performance data are summarized in Supplemental Table 2 (see annex, page 188-189). We aimed at 3% Sy/x limits (= measure for within-run imprecision) for the enzymes, but expanded them to 5% for ALT and AST because of the low concentration of both analytes in the reference interval (the study samples were from apparently healthy individuals; some samples were even at the limit of quantitation of certain assays).

1) Enzymes

Median Sy/x (%) values were generally in the order of 1.5% for ALP, 4% for ALT, and 2-2.5% for AST, GGT and LDH. ALT limits were exceeded by AU (5%), Vitros (8.1%), and Advia (5.9%). AST limits were exceeded by Vista (5.4%) while GGT limits were exceeded by DxC (4.6%).

At the mean concentration, peer CV limits for ALP were only exceeded by AU (5.1%). For ALT, the peer CV was generally high (median ~9%) and all but three assays exceeded the limits. Peer CV limits for AST were exceeded by Cobas (5.4%), and Vista (7.8%). Peer CV values for GGT were somewhat high (median ~7%) and GGT limits were exceeded by Modular (8.3%), Advia (8.8%), and Vista (8.1%). For LDH reverse the limits were exceeded by AU (8.8%). Peer CV limits at low concentration were exceeded frequently, with many additional violations as compared to the ones at mean concentration. However, violations were distributed among analytes and test systems so that individual listing is not useful. The data indicate a general need for improving peer comparability for enzymes at low concentration levels. The number of peer CV violations at high concentration is similar to the one at mean concentration, however, with a slight change in violation pattern. Limits for ALP were only exceeded by Advia (5.6%). For ALT, all assays (except Cobas + PP) exceeded the limits. Limits for AST were only exceeded by AU + PP (5.2%). Also for GGT only AU violated the limits (9.4%). DxC exceeded the limits for LDH forward (11.8%) and Advia for LDH reverse (7.4%). The observations for peer TE, generally, follow the ones for peer CV.

Most problems were observed for ALT and GGT. LDH forward and reverse peer performance was generally good.

2) Ion selective electrode assays (chloride, potassium, sodium)

Median Sy/x (%) values were generally in the order of 0.5% for all ion selective electrode (ISE) assays (chloride, potassium, sodium) and none exceeded the 1% limit. Median peer CVs were 1% for chloride, 1.2% for potassium, and 0.8% for sodium. Peer CV limits were only exceeded for Modular potassium (2.2%) and Modular sodium (1.1%).

At the low concentration range peer CV limits for chloride were exceeded by DxC (2.1%), Modular (1.9%), Advia (1.7%), and Vista (1.8%). In the high concentration range, limits were exceeded borderline for sodium (AU 1%, Cobas 1%, Modular 1.2%, and Vitros 1.1%).

Median peer TE values were 2.2% for chloride, 2.7% for potassium, and 1.6% for sodium. Peer TE limits were only exceeded for Modular potassium (4.5%).

B) AMTM/RMP-corrected AMTM performance

The assay performance for each analyte is discussed extensively in the manuscript on pages 171-183.

Figure 3 shows a graphical overview of the assay %-difference at the mean concentration. The bias limits (based on the biological variation model; see Table 2) were exceeded for ALP by only the DxC assay (-14.8%); for ALT, by all assays with PP activation but two (Vista and Vitros) and by all without PP (range: -25 to -40%); for AST by one assay with PP (Vista: -26.5%) and two without PP (Architect: -18.6%; Cobas: -18.8%); for GGT by four assays (biases ~ 15%, even -28% for the DxC) and two were borderline within the limits (Advia: -12.7%; Vista: 12.7%); for LDH, the forward reaction, by only one assay (DxC: -17.5%), for the reverse reaction, the most deviating assays (Vitros and Modular) differed by ~30%. For electrolytes, the assay comparability was good, except for chloride, particularly the Roche Cobas assay (-3.6%).

Table 5 shows similar information in numerical form. Bias differences are tabulated at low, mid (mean), and high concentration. Differences that exceed the limits (this time, based on state-of-the-art performance) are underlined.

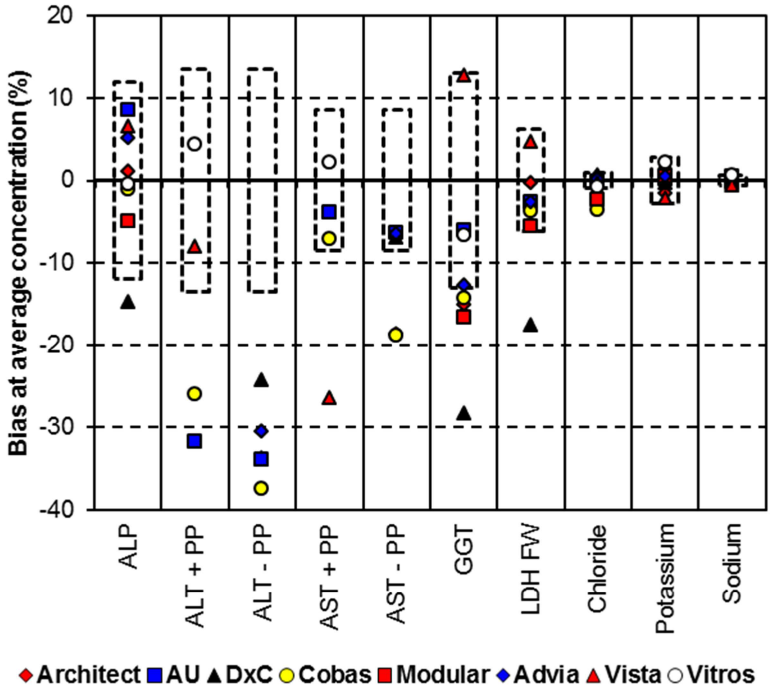


Figure 3. Assay %-difference at the mean concentration for ALT (26.0 U/L), AST (25.5 U/L) both with and without PP, GGT (25.9 U/L), and LDH forward (174.7 U/L) versus RMP target values, and for ALP (68.7 U/L), chloride (104.7 mmol/L), potassium (4.32 mmol/L) and sodium (140.1 mmol/L) versus AMTM target values. Note, LDH reverse data are not shown because of the high uncertainty of the AMTM. The black-broken boxes indicate the bias limits based on desirable biological variation.

Table 5. AMTM/RMP-corrected AMTM bias estimates (at low, mid, and high concentration) for each assay.

Bias (%)	Limits	Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros
ALP	10.2	1	7.9	-14.4	-1.9	-5.2	3.5	6.8	1
		1.1	8.5	-14.8	-1.1	-4.9	5.2	6.5	-0.5
		1.3	9	-14.9	-1.2	-4.8	3.9	6.4	-1.5
ALT + PP	9.5	N.A.	-39.8	N.A.	-29	N.A.	N.A.	15.1	50.6
		N.A.	-31.7	N.A.	-26	N.A.	N.A.	-8	4.4
		N.A.	-28.7	N.A.	-25.4	N.A.	N.A.	-12.8	-12.6
ALT - PP	9.5	-41	-38.3	-8.6	-39.2	N.A.	-24.2	N.A.	N.A.
		-33.8	-33.9	-24.2	-37.4	N.A.	-30.4	N.A.	N.A.
		-31	-31.2	-30	-36.3	N.A.	-29.9	N.A.	N.A.
AST + PP	7	N.A.	-9.5	N.A.	-9.2	N.A.	N.A.	-40.5	-0.5
		N.A.	-3.9	N.A.	-7.2	N.A.	N.A.	-26.5	2.2
		N.A.	-2.7	N.A.	-6.2	N.A.	N.A.	-18.5	2.5
AST - PP	7	-26	-11.3	-0.5	-21.7	N.A.	-2.1	N.A.	N.A.
		-18.6	-6.4	-7	-18.8	N.A.	-6.6	N.A.	N.A.
		-14.3	-3.2	-10.8	-16.2	N.A.	-5.1	N.A.	N.A.
GGT	9.5	-19	-8.9	-31	-18.4	-20.8	-20.3	16.3	-5.8
		-15.1	-6	-28.3	-14.3	-16.7	-12.7	12.7	-6.7
		-12	-5.7	-30.5	-11.1	-12.3	-7	9.7	-7.4
LDH FW	7	-0.4	-1.4	-10.4	-2.9	-5.2	-1.5	5.3	N.A.
		-0.3	-2.7	-17.5	-3.8	-5.6	-2.7	4.7	N.A.
		-0.2	-3.5	-14.5	-4.5	-5.8	-4.7	4.2	N.A.
LDH RV	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
		N.A.	-3.6	13.5	-10.4	-12.6	-6	N.A.	19.2
		N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Chloride	1.9	0.5	-0.4	-0.2	-3.7	-2.8	-0.2	-0.4	-0.9
		0.4	-0.2	0.6	-3.6	-2.3	0.2	-0.3	-0.7
		0.3	-0.2	0.2	-3.6	-2.5	0	-0.9	-0.6
Potassium	3	-1.6	0.2	-0.6	1.1	-0.1	0.7	-2	2.3
		-1.6	0.1	-0.3	1.1	0.4	0.4	-2.2	2.2
		-1.5	-0.2	0.3	1.2	0.1	0.4	-2.5	2.2
Sodium	1.4	-0.7	-0.1	-0.1	0.1	0	0.6	-0.2	0.5
		-0.5	-0.2	-0.3	0.2	0.4	0.3	-0.6	0.6
		-0.4	0	-0.5	0.4	0.4	0.5	-0.9	0.7

Chloride AMTM

The chloride results were characterized by good comparability of the assays, except the ones from Roche (Cobas and Modular). The latter ones were considerably lower than the others. We decided, therefore, to verify the results with the RMP procedure. Because we had no samples of the survey left, we let measure ten other samples in three Cobas routine laboratories (with two ISE units each, n = 6) that took part in the study and with the RMP (see Table 6 below).

The Cobas results were, on average, 4.3% (confidence interval 0.6%) lower than the RMP results. Similar to that, the Cobas results were ~3.6% lower than the AMTM without inclusion of the results by the Roche assays (which is visualized in the scatterplot of Figure 4). This let us conclude that the Roche ISE results were significantly biased and, therefore, were to be excluded them from the calculation of the AMTM. The RMP results, indirectly, support the validity of the AMTM.

Table 6. Chloride measurements

RMP (mmol/L)	Cobas Mean (mmol/L)	% difference
105.9	101.0	-4.7
112.4	107.0	-4.8
108.0	103.4	-4.3
106.8	101.8	-4.7
103.3	100.5	-2.7
98.0	94.5	-3.6
104.2	100.5	-3.6
103.6	97.1	-6.3
105.7	101.0	-4.4
101.1	97.3	-3.8
Average difference:		**-4.3** (0.6)

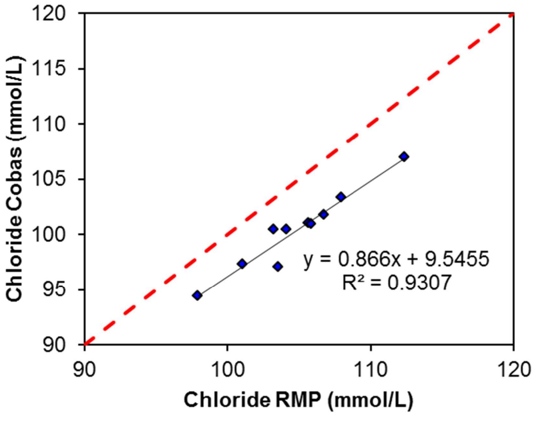


Figure 4. Scatter plot for chloride: RMP versus Cobas values.

Laboratory performance

Table 7 describes the combined imprecision (AMTM/RMP-corrected AMTM Sy/x) and bias (versus AMTM/RMP-corrected AMTM target) of all participating laboratories, including those which do not have a peer group. These values result from the combined effects of laboratories and assays. The differences between the laboratories are particularly influenced by assay bias.

One observes huge differences between the laboratories for all estimates. Maximum Sy/x values exceeding two times the AMTM/RMP-corrected AMTM limits (Table 7) are indicated by red shading. Very high Sy/x values are observed for enzymes (in particularly, for ALT, AST and GGT) and borderline for chloride.

Maximum absolute biases >15% (red shading) are observed for all enzymes in each concentration range. This leads frequently to differences >30% (blue shading) for the most deviating laboratories (= Diff 1) and even still for the third most deviating laboratories (= Diff 3). These differences mainly reflect the assay biases visualized in Figure 3. Maximum absolute biases >5% for chloride and potassium, and >2.5% for sodium (red shading) with the accompanying differences of >10% and >5% (blue shading) respectively, occur occasionally. However, most violations are borderline. A more graphical presentation of the complete distribution of laboratory averages, AMTM/RMP biases and AMTM/RMP Sy/x values can be found in the annex (Supplemental Figure 2).

A) Peer FAILs versus Peer Limits

Laboratories with poorer performance were identified by the frequency of peer failures (n ≥ 5 failures). 10% of the laboratories (n = 12) accounted for 52% FAIL decisions based on peer limits. The respective laboratories are advised to revisit their practices and/or consult the manufacturer for assistance.

B) Peer FAILs versus Fixed Limits

Laboratories which may profit from an improvement in performance were identified by the frequency of peer failures (n ≥ 6 failures) versus fixed limits. 11% of the laboratories (n = 14) accounted for 36% FAIL decisions based on fixed limits. The respective laboratories may see these results as an incentive for more rigorous QC.

Table 7. Overview of the AMTM/RMP-corrected AMTM of all laboratories

	ALP	ALT + PP	ALT - PP	AST + PP	AST - PP	GGT	LDH FW	Chloride	Potassium	Sodium
AMTM/RMP-corrected AMTM Sy/x										
Median	1.5	5.7	4.5	3.5	2.8	2.6	2.0	0.6	0.7	0.5
Min	0.7	2.3	1.7	2.1	1.4	1.0	1.1	0.2	0.1	0.2
Max	5.2	15.0	10.6	15.7	6.8	20.5	4.2	2.1	1.4	1.0
AMTM/RMP-corrected AMTM Bias										
Min	-19.5	-46.4	-54.7	-33.2	-23.8	-54.8	-20.1	-5.7	-5.2	-2.2
Max	18.5	28.2	-2.7	7.1	10.4	29.5	10.8	2.2	4.3	2.9
Diff 1	33.8	74.1	36.1	38.0	29.7	84.3	30.8	7.9	9.5	5.1
Diff 2	32.5	65.8	31.1	36.4	17.5	52.7	25.0	7.2	8.8	4.6
Diff 3	30.5	56.4	27.1	32.9	17.4	49.6	24.3	7.0	8.6	4.2
AMTM/RMP-corrected AMTM Bias Low										
Min	-19.0	-68.7	-67.4	-48.4	-32.1	-74.3	-19.1	-7.8	-5.5	-2.1
Max	22.2	112.6	26.4	9.6	21.2	52.1	12.4	2.0	4.6	2.8
Diff 1	37.3	174.6	69.8	57.8	45.6	126.4	31.5	9.8	10.0	4.9
Diff 2	32.2	147.3	47.8	54.9	31.9	75.9	28.1	8.1	8.9	4.5
Diff 3	31.2	122.2	42.4	50.5	29.8	60.2	25.9	7.8	8.8	4.5
AMTM/RMP-corrected AMTM Bias High										
Min	-19.9	-40.7	-51.2	-24.6	-22.0	-39.4	-20.8	-5.8	-4.8	-2.3
Max	18.0	8.4	-11.1	9.2	4.1	14.6	9.8	2.3	4.5	2.9
Diff 1	33.8	49.1	26.9	30.1	20.7	54.0	30.6	8.1	9.3	5.2
Diff 2	30.6	34.7	25.8	28.2	16.3	46.2	26.4	7.2	8.5	4.9
Diff 3	30.2	33.1	24.1	26.1	15.6	44.2	24.9	6.6	8.0	4.6

Note: LDH reverse is not mentioned since its AMTM is highly uncertain.

Study outcomes - laboratory communication: report and laboratory guide

Participating laboratories received two different Word documents and analyte-specific Excel files. One document, the general report, addressed all aspects of the survey (both laboratory and assay performance). The other document, the laboratory guide for data interpretation, assisted participants in interpreting the Excel files.

These files were anonymously coded. To interpret their data, participants could open the Excel report files and type in cell “A1” their laboratory code (and/or “spin” until their code appears). Each file contained four pages. Interpretation of the data was started with the investigation of the FAIL decisions in the “Overview” page 4 (see annex page 197). With this short summary the participants could easily see how their laboratory and how their assay performs versus peer group and AMTM/RMP-corrected AMTM targets. The PASS/FAIL decisions on this page were based on an interplay between the laboratory’s and assay’s performance. This combined-target approach was applied so laboratories weren’t, for example, blamed for bad quality when this was caused by poor assay performance (8). At the bottom of page 4 some recommendations were formulated of a general nature that should be “translated” into the laboratory and regulatory environment. For example, deviation of a laboratory from its peer group is a signal for the respective laboratory to contact the manufacturer, whilst violation of a manufacturer’s assay versus biologically derived bias limits might be a signal for the manufacturer for improving the calibration stability or the assay standardization or harmonization.

Afterwards, data could be investigated in more detail in the scatter-, %-difference, and %-residuals plots (pages 1 to 2; example annex page 196). Page 1 showed how the laboratory performs in comparison with its peer, and page 2 showed the comparison with the AMTM or RMP-corrected AMTM. Page 3 showed the distribution of the participants for the average analyte values, and histograms for the bias-% and Sy/x-% compared to the RMP-corrected AMTM. An overview of these distributions can be found in the annex (Supplemental Figure 2). Participants were recommended to “Spin” through the whole dataset to get an impression of the performance of other laboratories.

5. Conclusions on the Master Comparison 2014 survey

The design of our recent EQA study was well suited for monitoring the between-assay comparability and/or traceability of assays for five enzymes and three electrolytes. Peer performance (within-run imprecision and peer group variability) and assay comparability were generally good for all electrolytes. An exception was the negative bias of the Roche assays for chloride.

For enzyme assays both manufacturers and laboratories still have a lot to do to accomplish trueness of measurement. Peer performance (within-run imprecision and peer group variability) strongly depended on the enzyme measured. ALT assays were particularly problematic, possibly due to the nature of the samples (concentrations typically lie in the reference interval, whilst the higher concentrations are considered to be more clinically important). AMTM/RMP-corrected AMTM Sy/x values were similar to the peer estimates, indicating little influence of random sample-related effects. Comparability, or bias versus the RMP-corrected AMTM procedure, strongly depended on the enzyme measured.

In general, the conclusion for this Master Comparison survey is similar as its predecessors: “As good as conventional proficiency testing may be, we can do better” (20). The data show the need (i) to phase out the non-IFCC assay variants; (ii) to improve the quality, comparability, and standardization of the ALT assays; (iii) to improve the comparability of the GGT and LDH assays; (iv) to improve peer comparability at low concentration levels; (v) to aim for single-manufacturer efforts for improvement of comparability for nearly all enzymes studied.

References

1. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067-75.
2. Braga F, Panteghini M. Verification of in vitro medical diagnostics (IVD) metrological traceability: Responsibilities and strategies. Clin Chim Acta 2014;432:55-61.
3. Sciacovelli L, Secchiero S, Zardo L, Zaninotto M, Plebani M. External quality assessment: an effective tool for clinical governance in laboratory medicine. Clin Chem Lab Med 2006;44:740-9.
4. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232-5.
5. Libeer JC, Baadenhuijsen H, Fraser CG, Petersen PH, Ricós C, Stöckl D, Thienpont L. Characterization and classification of external quality assessment schemes (EQA) according to objectives such as evaluation of method and participant bias and standard deviation. External Quality Assessment (EQA) Working Group A on Analytical Goals in Laboratory Medicine. Eur J Clin Chem Clin Biochem 1996;34:665-78.
6. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670-80.
7. 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. Scand J Clin Lab Invest. 2003;63:195-201.
8. Stöckl D, Thienpont LM. The combined target approach - a way out of the proficiency testing dilemma. Arch Pathol Lab Med 1994;118:775-6.
9. Stöckl D, Libeer JC, Reinauer H, Thienpont LM, De Leenheer AP. Accuracy-based assessment of proficiency testing results with serum from single donations: possibilities and limitations. Clin Chem. 1996;42:469-70.
10. Stepman HC, Tiikkainen U, Stöckl D, Vesper HW, Edwards SH, Laitinen H, Pelanti J, Thienpont LM; Participating Laboratories. Measurements for 8 common analytes in native sera identify inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855-63.
11. Van Houcke SK, Rustad P, Stepman HC, Kristensen GB, Stöckl D, Røraas TH et al. Calcium, magnesium, albumin, and total protein measurement in serum as assessed with 20 fresh-frozen single-donation sera. Clin Chem 2012;58:1597-9.
12. Stepman HC, Stöckl D, Acheme R, Sesini S, Mazziotta D, Thienpont LM. Status of serum-calcium and -albumin measurement in Argentina assessed in 300

representative laboratories with 20 fresh frozen single donation sera. Clin Chem Lab Med 2011;49:1829-36.

13. Grubbs F. Procedures for detecting outlying observations in samples. Technometrics 1969,11:1-21.
14. Schumann G, Bonora R, Férard G, et al. IFCC Primary Reference Procedures for the Measurement of Catalytic Activity Concentrations of Enzymes at 37°C. Clin Chem Lab Med 2002;40:734–8.
15. WEQAS - Maintaining Quality in Laboratory Medicine. http://www.weqas.co.uk/ (accessed 29/09/2014).
16. RELA - IFCC External Quality assessment scheme for Reference Laboratories in Laboratory Medicine. http://www.dgkl-rfb.de:81/ (accessed 29/09/2014).
17. Kramer U, Kress M, Reinauer H, Spannagl M, Kaiser P. Candidate reference measurement procedures for chloride, potassium, sodium, calcium, magnesium, and lithium by inductively coupled plasma (isotope dilution) sector field mass spectrometry (ICP-(ID) SFMS) in serum. Clin Lab 2013;59:1017-29.
18. Joint Committee for Traceability in Laboratory Medicine (JCTLM). Database of higher-order reference materials, measurement methods/procedures and services. http://www.bipm.org/jctlm/ (Accessed 29/09/2014).
19. Westgard QC. Desirable specifications for total error, imprecision, and bias, derived from biologic variation. http://www.westgard. com/biodatabase1.html (accessed 02/07/2014).
20. Horowitz GL. Assessing accuracy on the front lines: a pragmatic approach for single-donor proficiency testing. Clin Chem 2014;60:806-8.

Article 4. Reference measurement procedure corrected all method trimmed mean (letter to editor)

Reference measurement procedure corrected all method trimmed mean - The best of two worlds

Kenneth Goossens[1], Linda M Thienpont[1*]

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

Reference measurement procedures (RMPs) are a vital part of the metrological traceability chain (1). They have one essential feature: they can break the commutability barrier for trueness transfer from artificial calibrations solutions (e.g., ethanolic standards) to the matrix of patient samples. RMPs often apply instrumental analytical procedures (e.g., mass spectrometry), require extensive sample clean-up, and involve manual steps (e.g., RMPs for enzymes) (2). These features may make them vulnerable to increased measurement imprecision, low-through-put and high measurement costs. Particularly the latter is prohibitive for RMPs performing a significant number of replicate measurements to reduce the analytical random error component. In contrast, routine procedures generally are characterized by very low within-run measurement imprecision in the order of 1-2% and performance at relatively low cost, which favors a high number of measurements. From this perspective, method comparison studies between a RMP and several routine procedures have the potential of combining “the best of two worlds”, i.e., the high trueness provided by the RMP and the low dispersion of the all method trimmed mean (AMTM) inferred from the results by the routine procedures. This can be accomplished by correcting the bias of the method comparison AMTM on the basis of its relationship to the RMP, to result in the so-called RMP-corrected AMTM. This requires that first linear regression is performed between the RMP and AMTM measurement results for the samples of the method comparison study. Then the regression equation is used to calculate the RMP-corrected AMTM values. This approach requires that the AMTM is sufficiently reliable and has a reduced random error component. The latter is accomplished if the AMTM is calculated from the results of a sufficient number of assays, i.e., 6 to 8. It can indeed be assumed that under these conditions the random sample-related effects in the measurement results are cancelled out.

Here, we investigated the potential of using the above described RMP-corrected AMTM for the assessment of imprecision and random sample-related effects of individual routine procedures. The data we used are from a method comparison study using 20 single donation samples from apparently healthy volunteers. The samples were measured for γ-glutamyltransferase (GGT) in singlet on the platforms of 8 manufacturers by a representative number of laboratories and the IFCC RMP performed in triplicate (2). The platforms were the Abbott Architect (used by 21 laboratories), Beckman AU (n = 19), Beckman DxC (n = 11), Ortho Vitros (n = 19), Roche Cobas (n = 26), Roche Modular (n = 9), Siemens Advia (n = 11) and Siemens Vista (n = 8). The AMTM for each sample was calculated as mean of the 8 peer group means, after investigation for outlying assays (3). Performing a Grubbs test did not identify any outlying assay. Also no single outlying values (z-value > 4) were identified, which means that trimming in this specific case was not necessary. Linear regression of the AMTM and the RMP values gave the following equation (including the 95% CI of slope and intercept): AMTM = 0.924 (± 0.045) RMP - 0.816 (± 1.25) ($R^2 = 0.99$), hence the RMP-corrected AMTM was obtained as (AMTM + 0.816)/0.924. We want to clarify that, because the data below are a comparison between RMP and RMP-corrected AMTM, the absolute values for both are practically the same.

Figure 1 shows the %-difference plots of the GGT results obtained by 2 routine procedures installed on 2 different modern platforms compared to the RMP (left) and the RMP-corrected AMTM values (right) as target. Using the magnitude of the R^2-values as a criterion, the relationship of the procedures' differences to the respective targets was best described by a logarithmic equation. However, the logarithmic relation was only used for the purpose of fitting, and should not be interpreted as a causal connection between the two methods. Data distribution and R^2-values indicate the superiority of using RMP-corrected AMTM values as target over the RMP-values as such. Indeed, from visual inspection of the %-difference plots the decrease of the scatter around the regression line is obvious when comparing the left and right part of the figure. Also the increase of the R^2-values is spectacular, i.e., from 0.17 to 0.73 for one platform and from 0.34 to 0.94 for the other. The dispersion of the data for regression to the RMP values (left) can mainly be attributed to the uncertainty of the RMP (in the order of 2 - 10%). In consequence, under these circumstances, random error effects specifically related to routine procedure (from high analytical imprecision and/or random sample-related effects) cannot be uncovered due to the overriding imprecision component of the RMP. While the bias information is not relevant for the assessment of combined random error effects, it is conserved in the

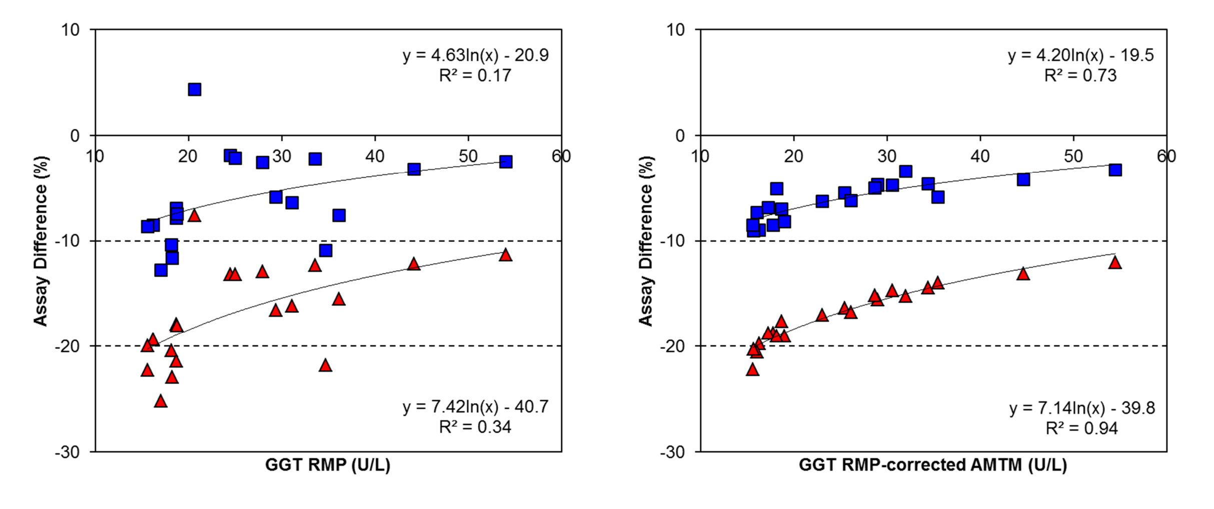


Figure 1. Demonstration of the effect of using RMP and RMP-corrected AMTM values on data dispersion around regression lines (logarithmic relationship); %-difference plots of results for γ-glutamyltransferase (GGT) by 2 different routine procedures compared to RMP values as such (left part) and RMP-corrected AMTM values (right part). The squared symbols and triangles used in the plots represent the differences of the respective routine procedures.

RMP-corrected AMTM approach because of the high quality of the regression equation relating the AMTM and the RMP values ($R^2 = 0.99$).

In conclusion, we demonstrated the superiority of using the RMP-corrected AMTM over RMP values as such when investigating causes of dispersion of results from routine procedures (imprecision and/or random sample-related effects) in a method comparison study. It is also an interesting approach in commutability studies in which one of the procedures is a RMP. These studies indeed may be limited by a high scatter of the results for the patient samples around the regression line due to the potentially higher measurement imprecision of the RMP, if not performed with sufficient replication because of too high costs.

References

1. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009; 55:1067–75.
2. Schumann G, Bonora R, Férard G, et al. IFCC Primary Reference Procedures for the Measurement of Catalytic Activity Concentrations of Enzymes at 37°C. Clin Chem Lab Med 2002;40:734–8.
3. Stepman HC, Tiikkainen U, Stöckl D, et al. Measurements for 8 common analytes in native sera identify inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855-63.

Article 5. Master Comparison 2014 survey

Trueness and comparability of widely used assays for 5 common enzymes and 3 electrolytes

Kenneth Goossens[1], Katleen Van Uytfanghe[1], Linda M Thienpont[1*], and Participating laboratories

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

For publication purposes, we had to reduce this article significantly. Nevertheless, in the interest of the reader of this thesis, the full-length publication is presented here.

Introduction

Current trends in healthcare policy are about reducing overall costs, while preserving quality of patient care. Examples of initiatives undertaken in this context are: promoting the development of uniform clinical practice guidelines for disease diagnosis and patient management, and/or algorithms for risk estimation from biochemical profiles (1, 2); providing evidence about the impact of treatment options on patient quality of life and survival (3); creating awareness of the general public about the importance of a healthy lifestyle and reacting towards early signals of health issues (4). Other initiatives are more compelling, such as the introduction of the electronic health record system (5). Laboratory medicine can substantially contribute in aiding these new developments, their implementation, and control, however, it is crucial that the discipline ensures that the data they generate are interchangeable and consistent. This is because guidelines, risk estimates and outcome related studies require aggregation of results from different research clinical trials or population studies, also because the informed patient of today often seeks for a second clinical opinion and might be highly confused by numerically different laboratory results from another source. Differences in laboratory results are caused by test principle and design, instrumentation, calibration material and commutability, etc. To eliminate the effect of these contributing factors, the concept of traceability in laboratory medicine has been developed (6). However, the success of this concept entirely depends on proper implementation by the stakeholders interested in providing and using standardized and interoperable laboratory medicine data (7). To fulfill this promise, initiatives likely to be successful are, among others, those creating evidence of the currently achieved standardization status, and/or awareness of laboratories and manufacturers where there are still limitations. External quality assessment (EQA) has earned a well-deserved position in this regard (8-11). However, the information provided by a common EQA design might be limited because of the use of processed materials not exactly behaving like patient samples (12). In addition, the focus is sometimes too much on “complex” analytes in the belief that the commonly performed high volume tests pose little problems (13). We emphasized these limitations in a recent study on the standardization status of diagnostic assays for 8 clinical chemistry analytes with use of single-donations samples (14). The conclusion of the accompanying editorial was “As good as conventional proficiency testing may be, we can do better. We are indebted to these authors for shedding light on a problem we may have assumed we did not have and, more important, for providing a powerful tool to help us make things better” (15). Therefore, we think there is a need to rebuild the system from scratch, meaning to start with the common high volume tests. In addition, it should be done from descriptive data that create reliable evidence and are in the same time appealing enough to be translated into action by laboratories and manufacturers. In that connection, we continued our initiative towards performing EQA surveys with a dedicated design to emphasize the trueness of assays and their comparability across manufacturers (14, 16, 17). This time we focused on widely used assays on modern platforms for analysis of 5 common enzymes and 3 electrolytes in serum.

Materials and methods

Study design and samples

Our EQA design consisted of the use of a panel of 20 fresh frozen, single donation serum samples (from Solomon Park Research Laboratories) [for a detailed description of the protocol for sample collection, storage, and handling we refer to (14)]. The samples were shipped on dry ice with the request to measure them for 8 analytes, i.e., alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), lactate dehydrogenase (LDH), gamma-glutamyl transferase (GGT), potassium, sodium and chloride. The measurements were done in singlet under within-run conditions. In total 126 laboratories were selected for participation to end up with 8 carefully controlled peer groups using the Abbott/Architect (n = 21), Beckman/AU (n = 19), Beckman/DxC (n = 11), Ortho/Vitros (n = 19), Roche/Cobas (n = 26), Roche/Modular (n = 9), Siemens/Advia (n = 12), and Siemens/Vista (n = 8) platforms. One loner participated with a Siemens/Xpand (n = 1) system. Also the manufacturers participated in the study with a minimum of 3 systems in their central application laboratories (data not shown nor discussed here). In spite of the selection of 8 peer groups, Table 1 shows that for certain enzymes quite often different methods are in use on one and the same platform, i.e., for ALP, the Deutsche Gesellschaft für Klinische Chemie (DGKC) versus International Federation of Clinical Chemistry (IFCC) method, for ALT/AST, with or without pyridoxal phosphate (PP) activation, for LDH, the forward and reverse reaction. The peer group sizes typically were at least 7; if a peer group had less than 7 participants, its estimates were only calculated provided the results were backed up by manufacturer data, or fitted in the generally expected picture of other assays (for example, LDH forward and reverse results). Certain peer groups could not be established because of too few participants, i.e., for ALP DGKC, ALT with PP (Architect, DxC, Modular, Advia), ALT without PP (Modular, Vista, Vitros), AST with PP (Architect, DxC, Modular, Advia), AST without PP (Modular, Vista, Vitros), LDH forward (Modular, Vitros), and LDH reverse (Architect, Vista).

Statistical data treatment

Reported results were all converted to Système International d'Unités (SI) Units. In case a laboratory reported use of a factor, its results were converted to the original value (without factor).

Single outlying results were identified by their z-value (>4) based on the median SD for the 20 samples of the respective peer group. Outlying laboratories (not included in the peer group) were identified by a 2-sided Grubbs test, based on 95% probability (18). Outlying assays (not included in the calculation of the all manufacturer trimmed mean (AMTM), see below) were also identified by a Grubbs test.

Target setting

Target values for data assessment were obtained from calculating the AMTM (in the case of ALP, LDH reverse, chloride, potassium and sodium) (14). For the enzymes AST, ALT, GGT and LDH forward the AMTM was corrected with the values set by the IFCC reference measurement procedures (RMP) (RMP-corrected AMTM) performed in the reference laboratory of WEQAS (Cardiff) (19-23). The RMP-corrected AMTM was calculated using the linear regression equation between the AMTM and the RMP values (AMTM = a RMP + b; RMP-corrected AMTM = (AMTM – b)/a). For chloride, 10 additional samples were measured with the Roche Cobas assay in 3 laboratories and by the ISO 17025/15195 reference laboratory of INSTAND (Düsseldorf) (24, 25). The above reference laboratories are ISO 17025/15195 accredited and take regularly part in the RELA - IFCC EQA scheme for Reference Laboratories in Laboratory Medicine (26).

Table 1 Assay fixed limits for AMTM/RMP-corrected AMTM estimates

	ALP	ALT*	AST*	GGT	LDH*	Cl	K	Na
Sy/x	3.0	5.0	5.0	3.0	3.0	1.0	1.0	1.0
Bias	10.2	9.5	7.0	9.5	7.0	1.9	3.0	1.4
Bias (biology)	11.9	13.5	8.5	13.1	6.3	0.9	2.8	0.6
Peer TE	15.1	17.7	15.2	14.4	11.9	3.5	4.6	3.0
AMTM/RMP-corrected AMTM unc.	5.2	2#	2#	2#	2#	0.4	1	0.4

*Limits are used for both ALT/AST with and without PP, but only for LDH forward
#Estimates refer to RMP-corrected AMTM targets

Specifications used for the assessment of the between-assay comparability and/or assay bias

The between-assay comparability and/or assay bias against the AMTM or RMP-corrected AMTM, as applicable, was estimated at 3 concentration levels (low, mid, and high). For assessment of the bias, specifications (Table 1) that accounted for the state-of-the-art performance were used. The bias limits were analyte-specific established from the peer group CV limits and the uncertainty of the AMTM or RMP-corrected AMTM target, as applicable. The former were set on the basis of the median peer group variation observed from the survey data (5% for ALP, AST and LDH; 7.5% for ALT and GGT; 1.5% for chloride, 2% for potassium, and 1% for sodium). Also bias specifications derived from biological variation were given, but for information, only (27, 28). Again the desirable limits were expanded for the uncertainty of the AMTM or RMP-corrected AMTM target.

Results

Outliers

Fifty-two single outliers out of 19480 results were identified (0.3%). Most of them were clerical errors. In addition, 20 laboratory outliers out of 975 tests were excluded (2.1%). These outliers gave no indication for assay quality, but were mainly related to peer groups with small variation (e.g., 4 outliers for chloride, 3 for sodium). No assay outliers were detected.

Assay comparability and bias

Figures 1 to 3 show the assay difference plots (%) versus the AMTM or RMP-corrected AMTM target values, with inclusion of the bias limits based on state-of-the-art performance (red dotted line) and biological variation (blue dotted line). Note, the here described data interpretation is against the above state-of-the-art limits.

Enzymes

For ALT with PP activation (Figure 1, top left) only the Vitros and Vista assays are at the mean concentration level within the ±9.5% bias limits compared to the RMP corrected-AMTM, however none of the assays meets the bias specifications over the whole concentration range, e.g., at the low concentration end the bias ranges from -40% to 50%, at the high end all assays are negatively biased (range from -29% to -13%). These biases generally results in poor between-assay comparability. Although the between-assay comparability is better for those without PP activation (Figure 1, top right), they again all are negatively biased over the whole concentration range (biases in

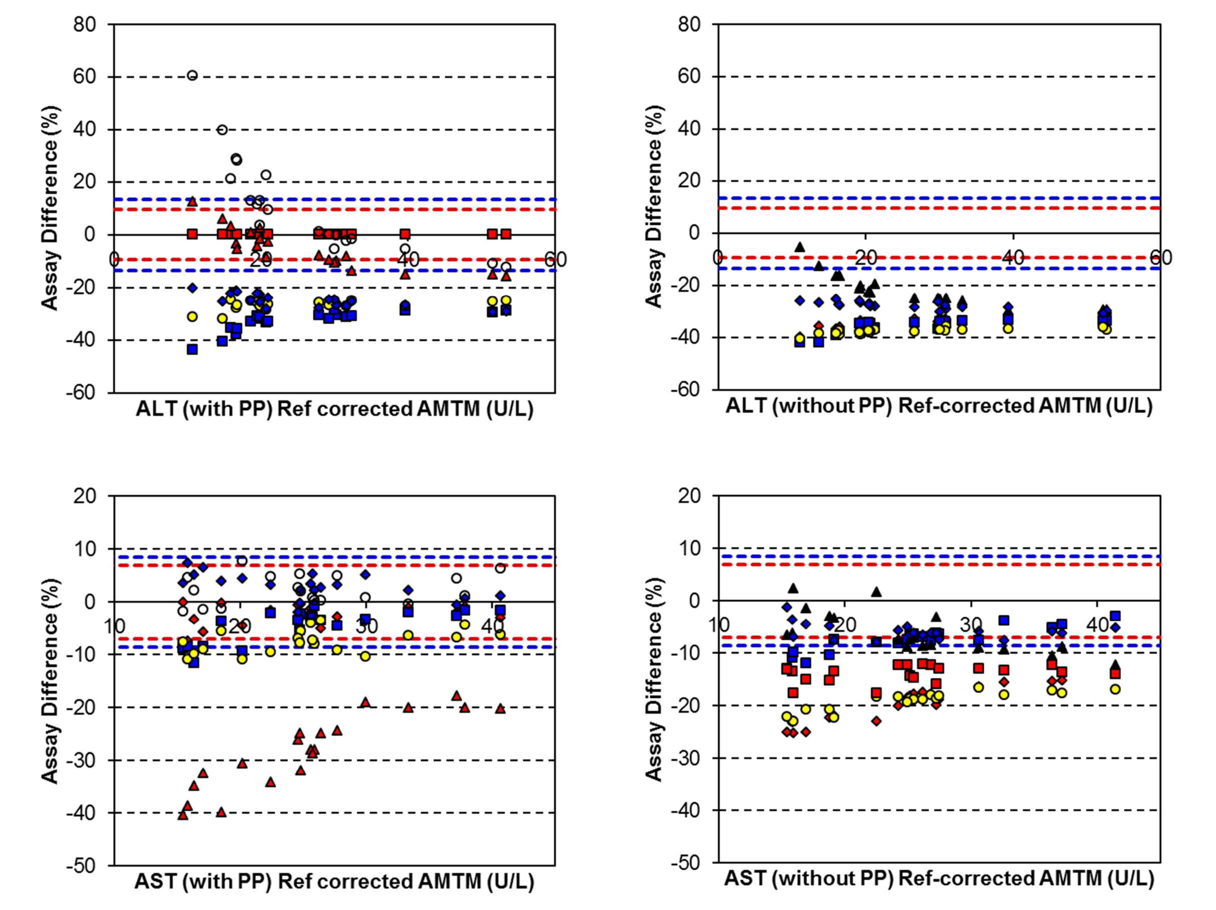


Figure 1. Assay %-difference for alanine transaminase (ALT), aspartate transaminase (AST) both with and without pyridoxal phosphate (PP) versus RMP-corrected AMTM target values, as applicable, for Abbott Architect (red diamond), Beckman AU (blue square), Beckman DxC (black triangle), Roche Cobas (yellow circle), Roche Modular (red square), Siemens Advia (blue diamond), Siemens Vista (red triangle), and Ortho Vitros (white circle). The red-broken bias limits are fixed limits based on state-of-the-art performance; the blue-broken limits are desirable bias limits from biological variation (both are listed in Supplemental Table S2) (25, 26).

the order of -30% to 40%) and exceed the limits. The Cobas assay without PP is ~10% lower than the variant with PP, while the two AU variants have a similar bias.

For AST with PP activation (Figure 1, bottom left), all assays but one (Vista) compare well and are within the ±7% bias limit (Cobas is borderline: -7.2%); their bias is independent of the concentration range. The Vista assay has a significantly negative bias in the high concentration range (-20%), which increases to -40% in the low range. The AST assays without PP activation (Figure 1, bottom right) are again all negatively biased over the whole concentration range, with the Architect and Cobas assays outside the bias limits (bias ~-20% at the mean concentration level.

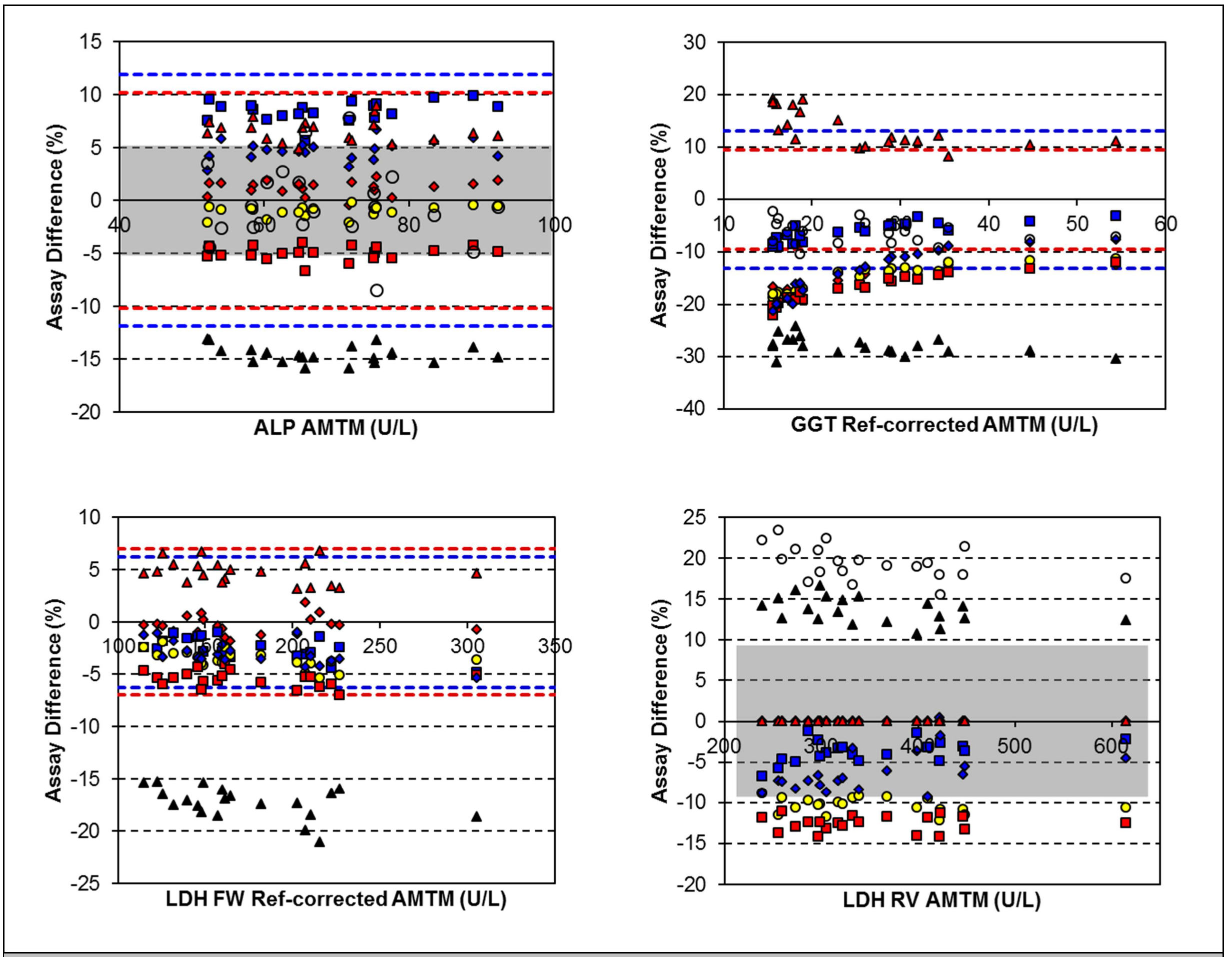


Figure 2. Assay %-difference for alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), lactate dehydrogenase (LDH) forward (FW) and reverse (RV) versus RMP-corrected AMTM or AMTM target values, as applicable, for Abbott Architect (red diamond), Beckman AU (blue square), Beckman DxC (black triangle), Roche Cobas (yellow circle), Roche Modular (red square), Siemens Advia (blue diamond), Siemens Vista (red triangle), and Ortho Vitros (white circle). The red and blue broken limits are the same as described for Fig. 1. The Grey zone in the LDH RV figure represent the uncertainty of the AMTM target.

and even higher at the low end). The Cobas assay without PP is in addition ~10-15% lower than the variant with PP. AU without PP is again similar to the variant with PP.

For ALP (Figure 2, top left), 7 of the 8 assays compare quite well with each other and meet the bias specifications of ~10.2%. Only the DxC assay exceeds the limits (-15%). The differences are proportional over the whole range.

The GGT assays (Figure 2, top right) show a significant between-assay discrepancy, with the 2 most deviating assays (Vista and DxC) differing by ~40-45%. Six out of eight assays (Architect, DxC, Cobas, Modular, Advia, Vista) are outside the bias limit of ±9.5%. The bias is nearly constant over the concentration range, apart from a moderate increase at the low end.

For the LDH forward reaction (Figure 2, bottom left) 6 out of 7 assays (Architect, AU, Cobas, Modular, Advia, Vista) are within the bias limit of ~7%. The Vista assay is positively biased, while the others negatively (with the AU most extremely), resulting in a difference of ~23% between the most discrepant assays.

Assays applying for LDH the reverse reaction (Figure 2, bottom right) give ~2 times higher values than those applying the forward one. Their AMTM is quite uncertain (~±10%), which is indicated in the figure with the grey zone. Therefore, those assays are not assessed versus the bias limit. Nevertheless, the maximum difference between the assays amounts to ~30% (Vitros +19.2%, Modular -12.6%). Note, although only a low number of laboratories measured LDH with the Modular forward and reverse reaction, the assay is shown because its performance is quite similar to the Cobas peer group.

Electrolytes

For chloride (Figure 3, bottom left), the 2 Roche assays (Cobas and Modular) are negatively biased as compared to the others (Modular: -2.3%, Cobas: -3.6% at the mean concentration) and exceed the ±1.9% limits, therefore, they were not included in the calculation of the AMTM. This was supported by the results obtained by a reference measurement procedure confirming the negative bias (results not shown here). The other assays even met the biological variation bias limit (±0.9%).

All potassium assays (Figure 3, top left) are within the state-of-the-art (±3%) and even the biological variation bias limits (±2.8%); the most deviating assays (Ortho and Vista) differ by ~4.5%.

For sodium, also all assays (Figure 3, top right) are within the state-of-the-art bias limits of ~1.4% and, even, nearly within the very tight limits (±0.6%) derived from biological variation.

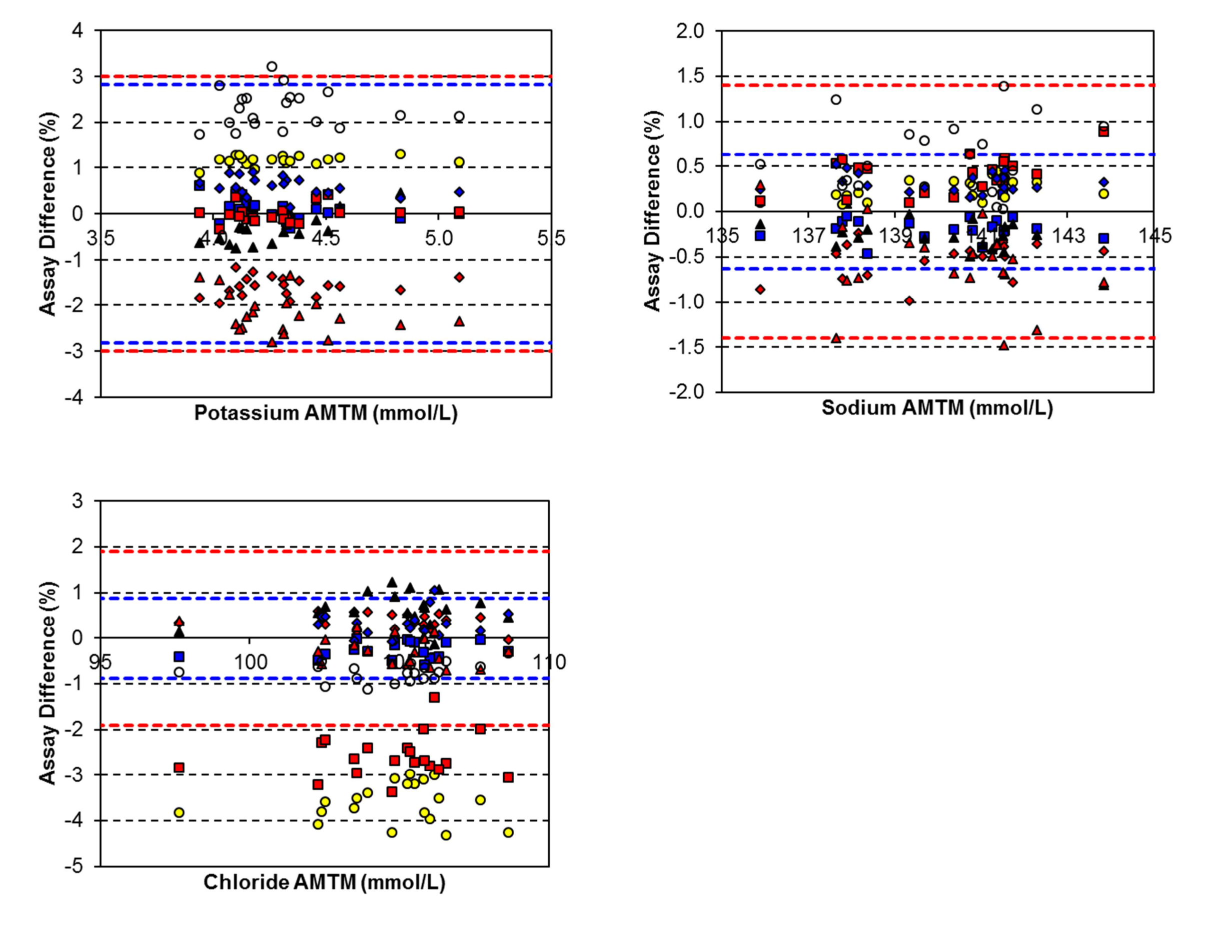


Figure 3. Assay %-difference for chloride, potassium, sodium versus AMTM target values, as applicable, for Abbott Architect (red diamond), Beckman AU (blue square), Beckman DxC (black triangle), Roche Cobas (yellow circle), Roche Modular (red square), Siemens Advia (blue diamond), Siemens Vista (red triangle), and Ortho Vitros (white circle). The red and blue broken limits are the same as described for Fig. 1.

Discussion

Targets and limits

As described, the limits used for assessment of bias respect the technical capabilities (state-of-the-art performance) and account for the uncertainty of the targets. Also the desirable bias limits inferred from biological variation are given, but for information only.

Samples were assigned with reference measurement procedure values for ALT, AST, GGT and LDH. These values were then used to correct the calculated and linearly related AMTM values. The resulting RMP-corrected AMTM target combines the high accuracy provided by the RMP and the low dispersion of the results on the basis of the AMTM. In addition, the trueness of the RMP values is conserved in the RMP-corrected AMTM approach, because of the high quality of the regression equations.

For chloride, measurement of 10 additional serum samples with the reference measurement procedure and the Roche Cobas assay in 3 laboratories indirectly proved the trueness of this study's AMTM target with exclusion of the Roche test results. For ALP, LDH reverse, sodium and potassium only the AMTM target was used. For sodium and potassium this was justified because of the low between-assay variation, and because the AMTM suits the purpose of this study. For ALP the AMTM uncertainty was relatively low, which justified the use of bias limits, whilst for LDH reverse the high uncertainty of the AMTM target excluded that any bias specification should be used.

Enzymes

ALP, ALT, AST, GGT, and LDH are key enzymes commonly requested for laboratory evaluation of liver - and bone disease, myocardial and pulmonary infarction (29, 30). Therefore, standardization of these assays received considerable attention in the past (19-22). Although it was strongly hoped that the combined effect of the availability of reference measurement procedures and the traceability requirement of European legislation would contribute to improving the between-assay comparability for enzymes, our study and others prove the opposite (31-33). We observed that several laboratories still use non-IFCC recommended methods (e.g., the DGKC methods for ALP and LDH). Also, the IFCC recommendation for PP activation in the assays for transaminases is not adopted, since the majority of laboratories belonging to certain peer groups still use the methods without PP. This points to the need for laboratories to take the responsibility of moving to the IFCC recommended methods, when available on the platform they use. Unfortunately, this is not always the case, e.g., on the Ortho Vitros system using the LDH reverse reaction. As far as concerns the peer groups using the IFCC recommended methods, our study strikingly revealed significant between-assay differences and/or biases. This holds particularly true for ALT and GGT, 2 assays for LDH, and 1 assay for ALP and AST. Because of these persistent problems, in the Netherlands it was decided to harmonize 7 common enzyme results through EQA, which showed quite successful (34). Naturally it would be more cost effective and less burdensome for the laboratories if standardization could be reached at the manufacturer level. Therefore, our study included the respective manufacturers (results not presented here) in the hope that the unique experimental design of our EQA study (20 single donations, carefully selected peer groups, manufacturer participation) will be an extra stimulus for further efforts toward global standardization of enzyme assays. With regard to the individual enzyme assays, our study showed that the AMTM trueness is good for AST. In contrast, the ALT assays exhibit, in general, a bias of -20% in the upper normal range. The transaminase assays without PP activation showed an expected negative bias of approximately -30% (ALT) and -10 to -15% (AST). The GGT group as a whole was slightly negatively biased, but as already indicated above, the major problem for this enzyme was the difference between the assays. The trueness of the LDH forward assay group was good. These observations demonstrates that, apart from the ALT case, the introduction of the IFCC reference measurement procedures positively impacted the global trueness of enzyme results, but that more effort is needed from the individual manufacturer to reach satisfactory trueness of their assay.

A potential weakness of our study might be seen in the fact that it used samples from apparently healthy volunteers, which resulted in enzyme concentrations in the normal range, where standardization might be considered less critical. However, it has been stressed that, at least in the upper normal range, standardization is as important, in particular, for longitudinal assessment (35). Also mildly elevated enzyme levels are becoming more important for diagnostic workup, and offer the potential for new applications (29, 30). But again, the full merit of traditional and new applications depends on the use of properly standardized and stable assays.

Electrolytes

The comparability between the electrolytes assays was generally quite satisfactory, giving evidence that all are traceable to the respective reference measurement procedures. Apart from the Roche chloride assays, most had biases to the AMTM even meeting (or almost) the biological limits. In view of this excellent standardization status, the most important parameter to control is the reagent/calibrator lot-to-lot variation in the manufacturing process and the stability of performance of the individual laboratory. As described, there are strong indications that the Cobas and Modular assays are effectively biased.

Conclusion

The design of our recent EQA study was again well suited for monitoring the between-assay comparability and/or trueness of assays for 5 enzymes and 3 electrolytes. The agreement between electrolyte assays, apart from one for chloride, was very good. For enzyme assays both manufacturers and laboratories still have a lot to do to accomplish trueness of measurement.

References

1. National Guideline Clearancehouse. http://www.guideline.gov/ (accessed 29/04/2014)
2. National Heart, Lung, and Blood Institute (NHLBI) Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack. http://cvdrisk.nhlbi.nih.gov/ (accessed 29/09/2014)
3. Arbor Research Collaborative for Health. Dialysis Outcomes and Practice Patterns Study. http://www.dopps.org (accessed 29/09/2014)
4. Healthy living. http://www.nlm.nih.gov/ medlineplus/healthyliving.html (accessed 29/09/2014)
5. The Official Web Site for the Medicare and Medicaid Electronic Health Records (EHR) Incentive Programs. http://www.cms. gov/Regulations-and-Guidance/Legislation /EHRIncentivePrograms/index.html (accessed 29/09/2014)
6. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067-75.
7. Braga F, Panteghini M. Verification of in vitro medical diagnostics (IVD) metrological traceability: Responsibilities and strategies. Clin Chim Acta 2014;432:55-61.
8. Sciacovelli L, Secchiero S, Zardo L, Zaninotto M, Plebani M. External quality assessment: an effective tool for clinical governance in laboratory medicine. Clin Chem Lab Med 2006;44:740-9.
9. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232-5.
10. Libeer JC, Baadenhuijsen H, Fraser CG, Petersen PH, Ricós C, Stöckl D, Thienpont L. Characterization and classification of external quality assessment schemes (EQA) according to objectives such as evaluation of method and participant bias and standard deviation. External Quality Assessment (EQA) Working Group A on Analytical Goals in Laboratory Medicine. Eur J Clin Chem Clin Biochem 1996;34:665-78.
11. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670-80.
12. Stöckl D, Thienpont LM. The combined target approach - a way out of the proficiency testing dilemma. Arch Pathol Lab Med 1994;118:775-6.
13. Sturgeon C. External quality assessment schemes for immunoassays. Methods Mol Biol 2013;1065:291-305.
14. Stepman HC, Tiikkainen U, Stöckl D, Vesper HW, Edwards SH, Laitinen H, Pelanti J, Thienpont LM; Participating Laboratories. Measurements for 8

common analytes in native sera identify inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855-63.

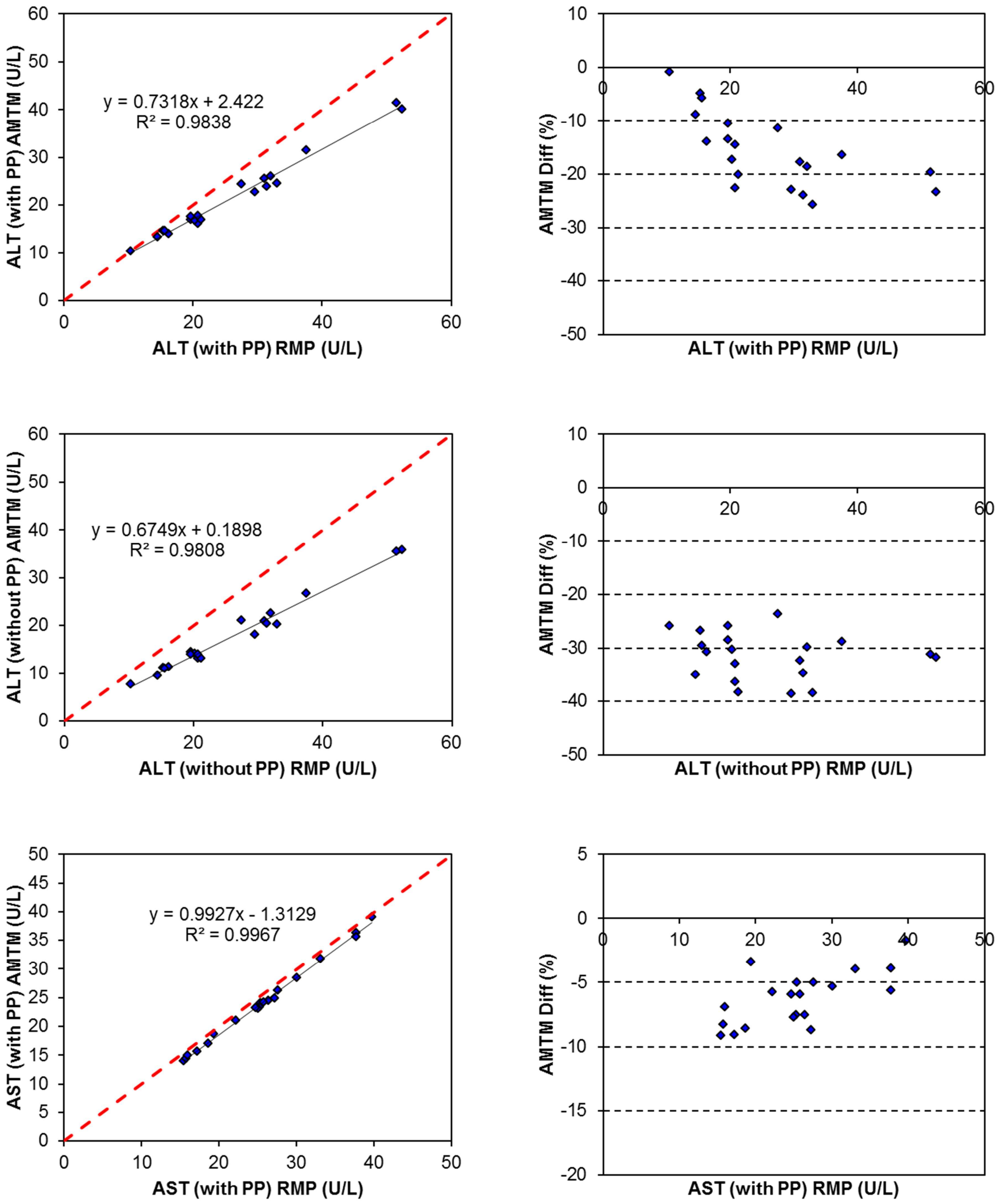
15. Horowitz GL. Assessing accuracy on the front lines: a pragmatic approach for single-donor proficiency testing. Clin Chem 2014;60:806-8.
16. Stepman HC, Stöckl D, Acheme R, Sesini S, Mazziotta D, Thienpont LM. Status of serum-calcium and -albumin measurement in Argentina assessed in 300 representative laboratories with 20 fresh frozen single donation sera. Clin Chem Lab Med 2011;49:1829-36.
17. Van Houcke SK, Rustad P, Stepman HC, Kristensen GB, Stöckl D, Røraas TH et al. Calcium, magnesium, albumin, and total protein measurement in serum as assessed with 20 fresh-frozen single-donation sera. Clin Chem 2012;58:1597-9.
18. Grubbs F. Procedures for detecting outlying observations in samples. Technometrics 1969,11:1-21.
19. Schumann G, Bonora R, Ceriotti F, Férard G, Ferrero CA, Franck PFH et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. Part 5. Reference procedure for the measurement of catalytic concentration of aspartate aminotransferase. Clin Chem Lab Med 2002;40:725-33.
20. Schumann G, Bonora R, Ceriotti F, Férard G, Franck PFH, Gella F-J et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. Part 4. Reference procedure for the measurement of catalytic concentration of alanine amino-transferase. Clin Chem Lab Med 2002;40:718-24.
21. Schumann G, Bonora R, Ceriotti F, Férard G, Ferrero CA, Franck PFH et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37C. Part 6. Reference procedure for the measurement of catalytic concentration of γ-glutamyltransferase. Clin Chem Lab Med 2002;40: 734-38.
22. Schumann G, Bonora R, Ceriotti F, Clerc-Renaud P, Ferrero CA, Férard G 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;40:643-48.
23. WEQAS - Maintaining Quality in Laboratory Medicine. http://www. weqas.co.uk/ (accessed 29/09/2014)
24. Kramer U, Kress M, Reinauer H, Spannagl M, Kaiser P. Candidate reference measurement procedures for chloride, potassium, sodium, calcium, magnesium, and lithium by inductively coupled plasma (isotope dilution) sector field mass spectrometry (ICP-

(ID) SFMS) in serum. Clin Lab 2013;59:1017-29.

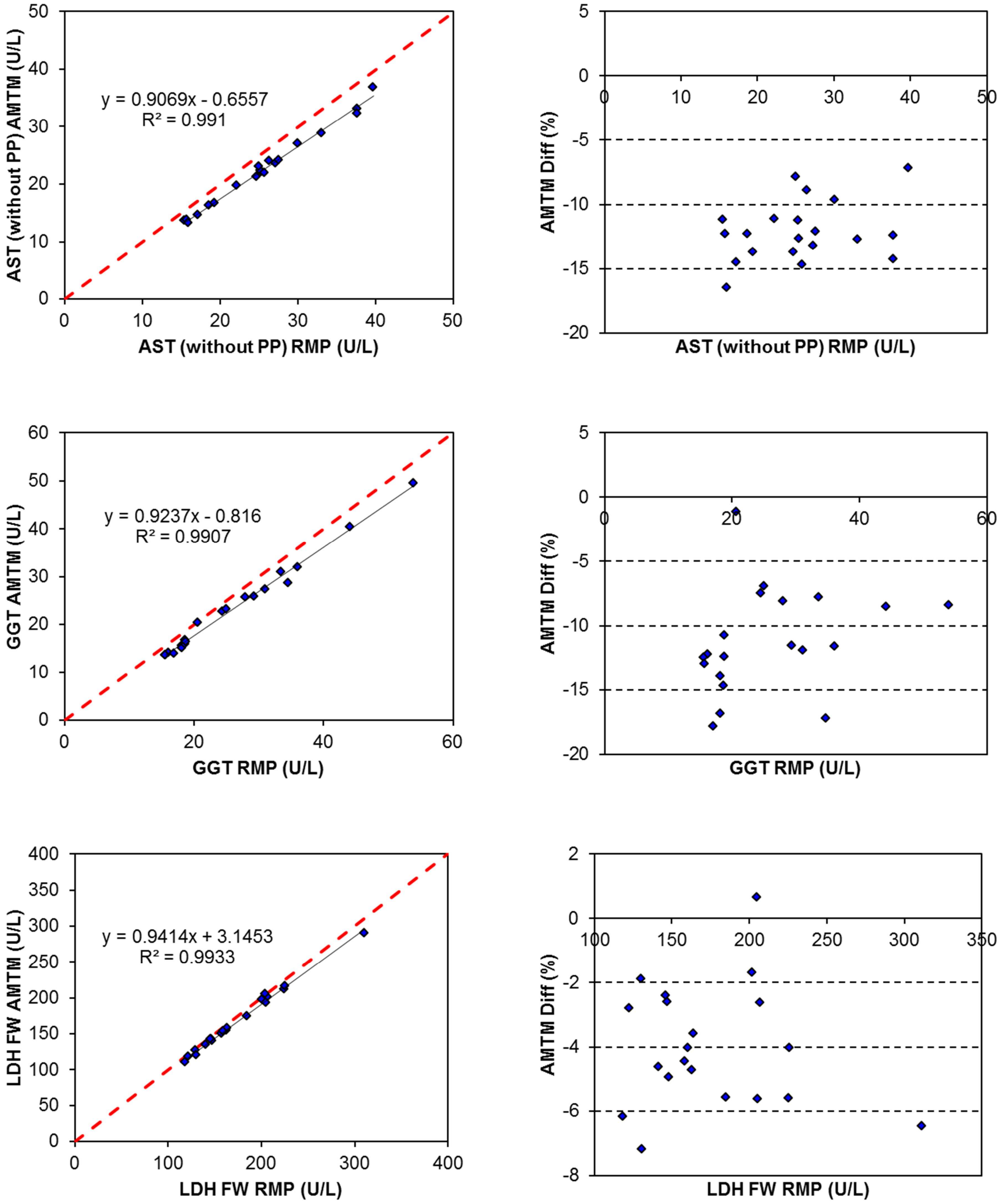
25. Joint Committee for Traceability in Laboratory Medicine (JCTLM). Database of higher-order reference materials, measurement methods/ procedures and services. http://www.bipm.org/jctlm/ (Accessed 29/09/2014).
26. RELA - IFCC External Quality assessment scheme for Reference Laboratories in Laboratory Medicine. http://www.dgkl-rfb.de:81/ (accessed 29/09/2014)
27. Fraser CG. Biological Variation: from principles to practice. Washington DC: AACC press; 2001.
28. Desirable specifications for total error, imprecision, and bias, derived from biologic variation. http://www.westgard.com/biodatabase1.html (accessed 02/072014).
29. Robert COH, Thomas RH. Causes and evaluation of mildly elevated liver transaminase levels. American Family Physician 2011; 48:1003-8.
30. Liu Z, Que S, Xu J, Peng T. Alanine aminotransferase – Old biomarker and new concept: a review. Int J Med Sci 2014; 11:925-35.
31. Infusino I, Schumann G, Ceriotti F, Panteghini M. Standardization in clinical enzymology: a challenge for the theory of metrological traceability. Clin Chem Lab Med 2010; 48: 301-307.
32. Ceriotti F, Henny J, Queraltó J, Ziyu S, Özarda Y, Chen B. Common reference intervals for aspartate amino-transferase (AST), alanine amino-transferase (ALT) and γ-glutamyl transferase (GGT) in serum; results from an IFCC multicenter study. Clin Chem Lab Med 2010; 48:1593-1601.
33. Panteghini M, Ceriotti F. Obtaining reference intervals traceable to reference measurement systems: is it possible, who is responsible, what is the strategy? Clin Chem Lab Med 2011;50:813-7.
34. Weykamp C, Franck P, Gunnewiek JK, de Jonge R, Kuypers A, van Loon D, Steigstra H, Cobbaert C. Harmonisation of seven common enzyme results through EQA. Clin Chem Lab Med 2014 May 29. pii: /j/cclm.ahead-of-print/cclm-2014-0458/cclm-2014-0458.xml. doi: 10. 1515/cclm-2014-0458. [Epub ahead of print].
35. Wu AHB. Standardization of assays for clinically important enzymes that have high biologic variation: what is all the fuss about? Clin Chem Lab Med 2010;48:299-300.

Annex to chapter 4

Regression equations and %-difference plots between the AMTM and the RMP procedure:



Supplemental Figure 1 (part 1). Regression equations and %-difference plots between the AMTM and the RMP for ALT and AST (both with and without PP), GGT, and LDH forward (FW).



Supplemental Figure 1 (part 2). Regression equations and %-difference plots between the AMTM and the RMP for ALT and AST (both with and without PP), GGT, and LDH forward (FW).

Peer group dependent laboratory limits:

Supplemental Table 1. Peer group-dependent laboratory limits

2 * Median Peer Sy/x (%)	Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros
ALP	2.5	2.7	4.3	1.9	1.2	2.2	4.3	3.0
ALT	NA	10.1	NA	5.4	NA	NA	6.4	16.2
ALT - PP	8.3	10.3	9.3	6.8	NA	12.7	NA	NA
AST	NA	4.6	NA	7.3	NA	NA	10.9	4.1
AST - PP	4.6	5.6	6.0	5.4	NA	6.0	NA	NA
GGT	3.8	3.7	9.1	3.9	4.2	5.0	5.9	5.4
LDH FW	4.3	3.5	3.9	3.3	NA	3.0	4.1	NA
LDH RV	NA	4.3	4.4	3.0	3.3	4.7	NA	5.4
Chloride	0.8	1.0	1.1	1.1	1.1	1.0	0.9	0.8
Potassium	1.3	1.1	1.3	1.0	1.3	1.2	1.5	1.5
Sodium	0.8	0.9	0.9	0.8	1.1	0.6	1.0	0.9
2 * Peer CV (%)	**Architect**	**AU**	**DxC**	**Cobas**	**Modular**	**Advia**	**Vista**	**Vitros**
ALP	5.2	9.5	5.5	5.2	3.7	11.2	9.8	9.5
ALT	NA	21.9	NA	15.9	NA	NA	10.3	24.6
ALT - PP	37.1	32.3	14.7	11.1	NA	34.7	NA	NA
AST	NA	5.6	NA	9.7	NA	NA	14.4	6.6
AST - PP	7.9	7.6	9.2	6.5	NA	5.8	NA	NA
GGT	9.9	9.1	7.8	13.5	16.2	16.0	16.2	13.9
LDH FW	4.2	6.6	5.7	4.3	NA	7.3	5.5	NA
LDH RV	NA	17.6	5.6	4.2	2.1	10.7	NA	4.5
Chloride	1.4	1.5	1.9	2.2	2.6	2.1	1.6	2.2
Potassium	3.7	1.7	2.1	3.0	5.0	2.3	0.9	2.5
Sodium	1.3	1.5	1.5	1.9	2.2	0.9	1.1	2.0
2 * Peer CV Low (%)	**Architect**	**AU**	**DxC**	**Cobas**	**Modular**	**Advia**	**Vista**	**Vitros**
ALP	6.0	10.3	5.7	7.4	3.8	9.4	12.3	11.4
ALT	NA	34.5	NA	41.1	NA	NA	17.5	38.1
ALT - PP	52.6	27.5	13.0	27.4	NA	56.7	NA	NA
AST	NA	9.9	NA	11.3	NA	NA	29.2	12.7
AST - PP	9.5	7.8	11.7	11.4	NA	21.6	NA	NA
GGT	10.6	10.0	42.0	14.0	18.8	27.6	32.5	22.8
LDH FW	5.3	10.3	30.7	5.1	NA	7.3	6.3	NA
LDH RV	NA	11.0	7.0	6.1	2.8	12.9	NA	9.8
Chloride	1.6	1.6	4.2	3.0	3.8	3.4	3.5	2.3
Potassium	4.0	2.0	2.5	3.2	6.6	2.3	1.1	2.6
Sodium	1.4	1.9	1.6	1.8	2.3	1.9	1.5	1.9

NA: not applicable

2 * Peer CV High (%)	Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros
ALP	4.9	8.9	5.5	7.1	3.7	11.2	8.2	8.5
ALT	NA	20.0	NA	9.4	NA	NA	23.1	16.8
ALT - PP	33.1	26.5	16.6	19.3	NA	24.5	NA	NA
AST	NA	10.4	NA	6.4	NA	NA	9.0	8.4
AST - PP	8.1	6.0	8.5	9.4	NA	8.9	NA	NA
GGT	10.6	18.8	11.3	13.4	15.0	9.5	4.7	7.6
LDH FW	5.6	8.8	23.5	4.7	NA	7.4	5.7	NA
LDH RV	NA	8.8	5.4	3.1	2.6	14.7	NA	8.3
CL	1.4	1.5	3.9	2.1	2.2	3.0	2.8	2.3
K	3.3	1.8	1.7	2.9	4.8	2.5	1.3	2.6
NA	1.5	2.0	1.5	2.0	2.5	1.6	1.0	2.1

Lab Peer TE (%) (1.645x)	Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros
ALP	9.3	15.9	11.2	9.0	6.3	18.5	17.3	16.1
ALT	NA	38.9	NA	27.0	NA	NA	19.7	47.5
ALT - PP	61.2	54.7	28.1	21.0	NA	59.6	NA	NA
AST	NA	11.7	NA	19.5	NA	NA	29.0	12.5
AST - PP	14.8	15.2	17.7	13.6	NA	13.5	NA	NA
GGT	17.1	15.8	19.4	22.6	27.0	27.1	27.8	24.1
LDH FW	9.7	12.1	11.2	8.8	NA	12.8	11.0	NA
LDH RV	NA	29.3	11.5	8.4	6.2	18.9	NA	11.4
CL	2.6	3.0	3.5	3.9	4.6	3.7	3.0	3.8
K	6.3	3.2	4.0	5.1	8.4	4.2	2.8	4.6
NA	2.5	2.8	2.8	3.3	4.0	1.7	2.4	3.5

NA: not applicable

Peer group performance data:

Supplemental Table 2. Peer group estimates for Sy/x, CV, and TE (values exceeding the limits are shaded in red; limits are listed in red in the left column)

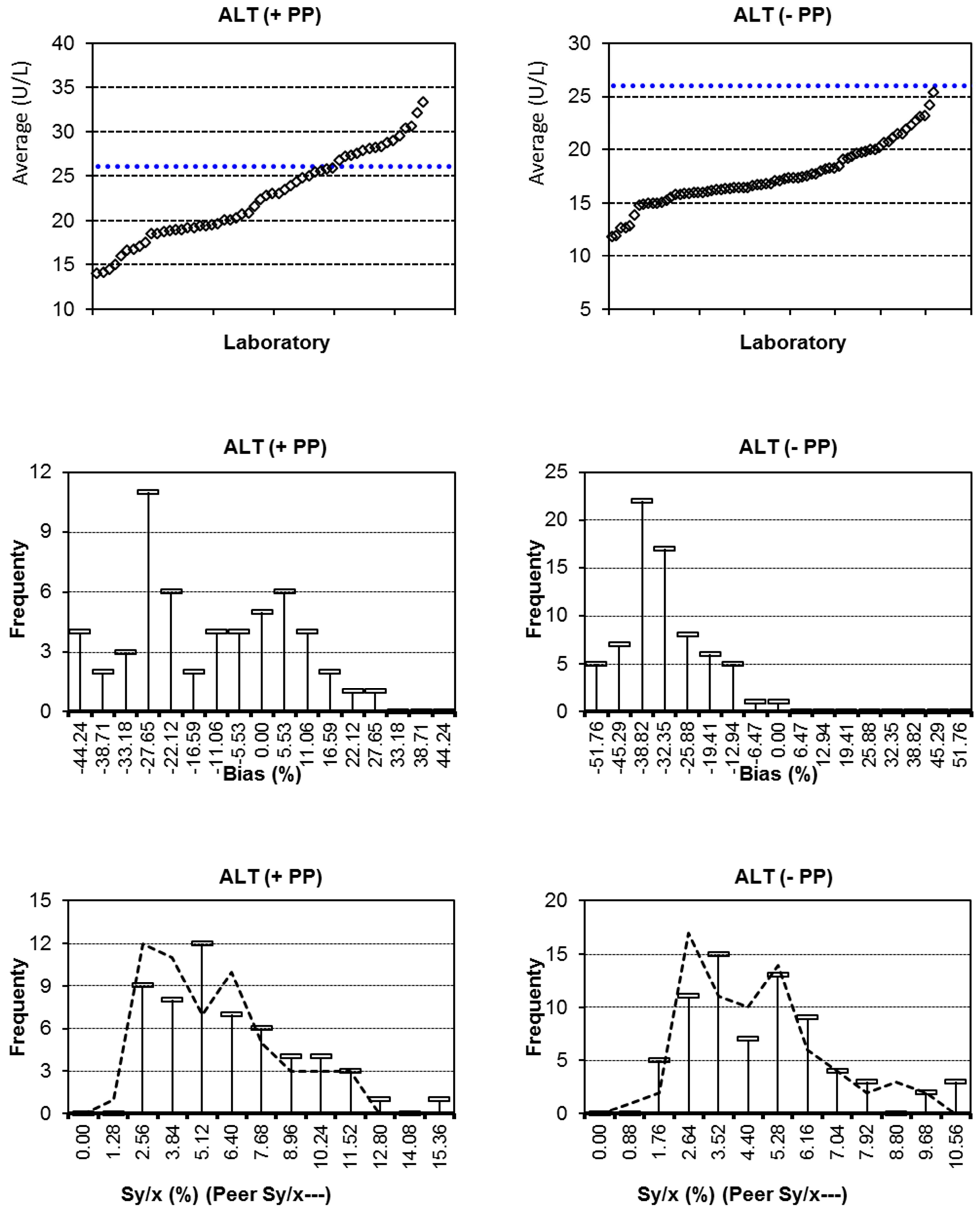
Median Peer Sy/x (%)		Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros	*Median*
3	ALP	1.3	1.3	2.1	0.9	0.5	1.1	2.2	1.5	*1.3*
5	ALT	NA	5.0	NA	2.9	NA	NA	3.2	8.1	*3.4*
5	ALT - PP	4.2	4.4	4.7	3.4	NA	5.9	NA	NA	*4.4*
5	AST	NA	2.3	NA	4.2	NA	NA	5.4	2.1	*2.3*
5	AST - PP	2.3	2.7	3.0	2.7	NA	2.5	NA	NA	*2.7*
3	GGT	1.9	1.8	4.6	1.9	2.2	2.4	3.0	2.7	*2.3*
3	LDH FW	2.2	1.7	2.0	1.7	NA	1.2	2.1	NA	*1.7*
3	LDH RV	NA	2.2	2.2	1.5	1.6	2.4	NA	2.7	*2.2*
1	CL	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.4	*0.5*
1	K	0.6	0.5	0.6	0.5	0.6	0.6	0.7	0.7	*0.6*
1	NA	0.4	0.4	0.4	0.4	0.5	0.3	0.5	0.5	*0.4*
	Median	1.3	1.8	2.1	1.7	1.1	2.2	2.1	1.8	
Peer CV (%)		**Architect**	**AU**	**DxC**	**Cobas**	**Modular**	**Advia**	**Vista**	**Vitros**	***Median***
5	ALP	2.6	5.1	2.8	2.6	1.9	2.0	4.9	4.8	*2.7*
7.5	ALT	NA	11.0	NA	8.7	NA	NA	5.2	12.3	*8.7*
7.5	ALT - PP	18.5	13.3	7.4	5.6	NA	9.5	NA	NA	*9.5*
5	AST	NA	2.8	NA	5.4	NA	NA	7.8	3.3	*3.1*
5	AST - PP	4.0	2.9	4.6	3.3	NA	3.2	NA	NA	*3.6*
7.5	GGT	5.0	4.2	3.9	6.7	8.3	8.8	8.1	7.0	*6.9*
5	LDH FW	2.1	3.3	2.9	2.2	NA	3.9	2.7	NA	*2.7*
5	LDH RV	NA	8.8	2.8	2.1	1.0	5.3	NA	2.3	*2.5*
1.5	CL	0.7	0.7	1.0	1.1	1.1	1.1	0.8	1.1	*1.0*
2	K	1.8	0.8	1.1	1.5	2.2	1.2	0.5	1.2	*1.2*
1	NA	0.7	0.7	0.8	0.9	1.1	0.4	0.5	1.0	*0.8*
	Median	2.1	3.3	2.8	2.6	1.8	2.7	3.8	2.8	
Peer CV Low (%)		**Architect**	**AU**	**DxC**	**Cobas**	**Modular**	**Advia**	**Vista**	**Vitros**	***Median***
5	ALP	3.0	5.2	2.9	3.7	1.9	4.7	6.2	5.7	*4.2*
7.5	ALT	NA	17.3	NA	20.6	NA	NA	8.7	19.0	*18.1*
7.5	ALT - PP	26.3	13.8	6.5	13.7	NA	30.2	NA	NA	*13.8*
5	AST	NA	4.9	NA	5.6	NA	NA	14.6	6.3	*6.0*
5	AST - PP	4.8	3.9	5.9	5.7	NA	10.8	NA	NA	*5.7*
7.5	GGT	5.3	5.0	21.0	7.0	9.4	13.8	16.3	11.4	*10.4*
5	LDH FW	2.7	5.1	15.4	2.6	NA	3.7	3.2	NA	*3.4*
5	LDH RV	NA	5.5	3.5	3.0	1.4	6.4	NA	4.9	*4.2*
1.5	CL	0.8	0.8	2.1	1.5	1.9	1.7	1.8	1.1	*1.6*
2	K	2.0	1.0	1.2	1.6	2.8	1.2	0.5	1.3	*1.3*
1	NA	0.7	1.0	0.8	0.9	1.2	1.0	0.7	1.0	*0.9*
	Median	2.8	5.0	3.5	3.7	1.9	4.7	4.7	5.3	
Peer CV high(%)		**Architect**	**AU**	**DxC**	**Cobas**	**Modular**	**Advia**	**Vista**	**Vitros**	***Median***
5	ALP	2.4	4.4	2.7	3.6	1.8	5.6	4.1	4.3	*3.8*
7.5	ALT	NA	10.0	NA	4.7	NA	NA	11.6	8.4	*9.2*
7.5	ALT - PP	16.6	13.3	8.3	9.6	NA	13.0	NA	NA	*13.0*
5	AST	NA	5.2	NA	3.2	NA	NA	4.5	4.2	*4.3*
5	AST - PP	4.0	3.0	4.2	4.7	NA	4.5	NA	NA	*4.2*
7.5	GGT	5.3	9.4	5.6	6.7	7.5	4.8	2.3	3.8	*5.5*
5	LDH FW	2.8	4.4	11.8	2.4	NA	3.7	2.9	NA	*3.3*
5	LDH RV	NA	4.4	2.7	1.5	1.3	7.4	NA	4.2	*3.4*
1.5	CL	0.7	0.8	2.0	1.1	1.1	1.5	1.4	1.1	*1.1*
2	K	1.7	0.9	0.9	1.5	2.1	1.3	0.7	1.3	*1.3*
1	NA	0.8	1.0	0.8	1.0	1.2	0.8	0.5	1.1	*0.9*
	Median	2.6	4.4	2.7	3.2	1.6	4.5	2.6	4.0	

NA: not applicable (low number of laboratories)

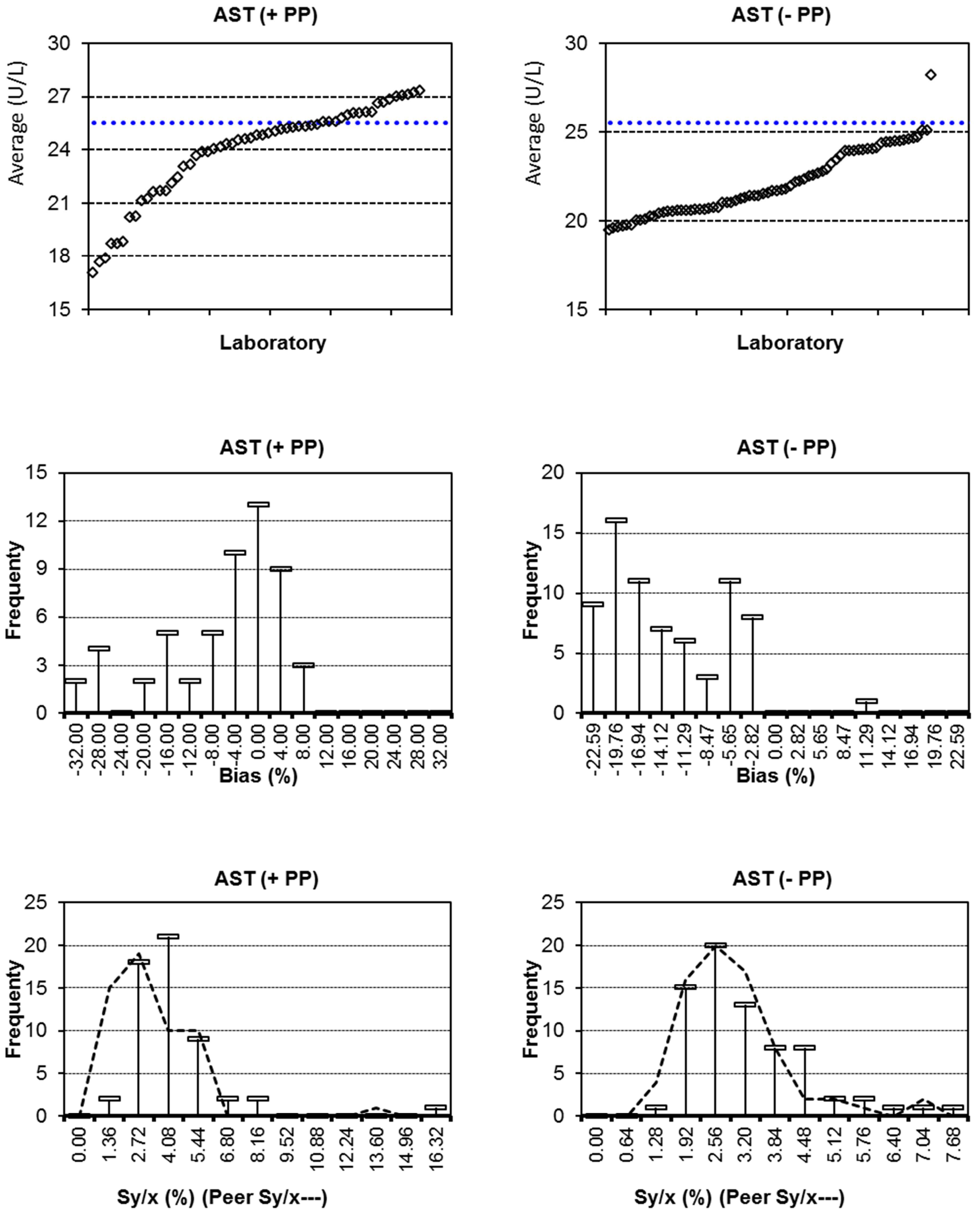
Peer TE (%)		Architect	AU	DxC	Cobas	Modular	Advia	Vista	Vitros	*Median*
11.4	ALP	5.6	10.3	6.8	5.5	3.8	4.4	10.5	9.8	*6.2*
17.7	ALT	NA	23.6	NA	18.0	NA	NA	11.9	28.9	*18.0*
17.7	ALT - PP	37.2	27.5	17.1	12.8	NA	22.0	NA	NA	*22.0*
13.9	AST	NA	7.1	NA	13.5	NA	NA	18.5	7.6	*7.4*
13.9	AST - PP	9.0	7.8	10.7	8.3	NA	8.0	NA	NA	*8.6*
15.8	GGT	10.4	9.0	11.8	13.7	16.9	17.9	16.9	14.6	*14.2*
11.4	LDH FW	5.9	7.4	6.8	5.3	NA	8.0	6.7	NA	*6.7*
11.4	LDH RV	NA	17.8	7.0	5.1	3.8	11.5	NA	6.9	*7.0*
3.5	CL	1.6	1.8	2.2	2.4	2.4	2.3	1.8	2.3	*2.2*
4.4	K	3.8	1.8	2.4	3.1	4.5	2.6	1.7	2.8	*2.7*
2.8	NA	1.5	1.7	1.7	2.0	2.4	1.0	1.4	2.2	*1.7*
	Median	5.6	7.8	6.8	5.5	4.2	7.8	8.6	7.3	

NA: not applicable (low number of laboratories)

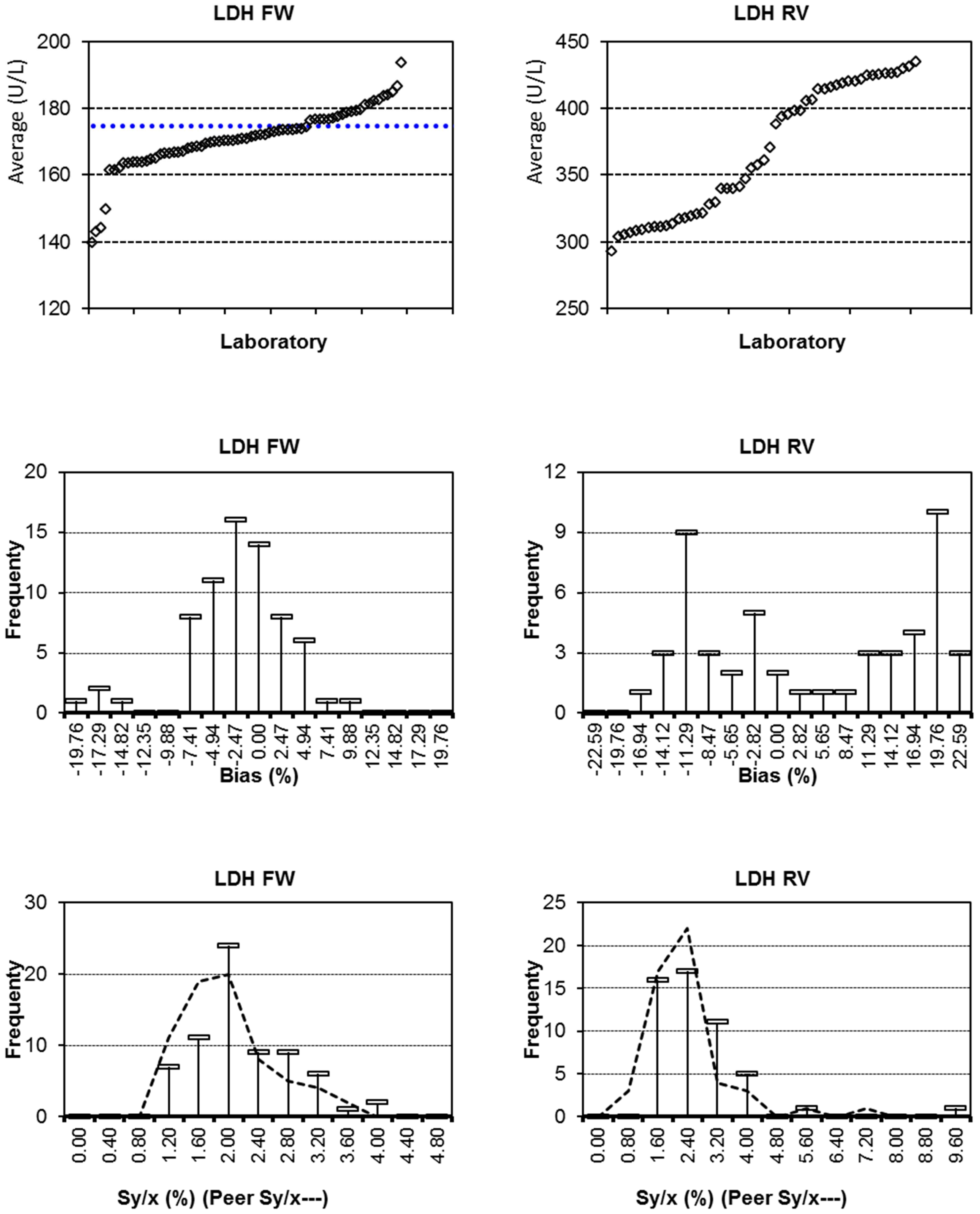
Graphical presentation of the complete distribution of laboratory averages, AMTM/RMP biases and AMTM/RMP Sy/x values:



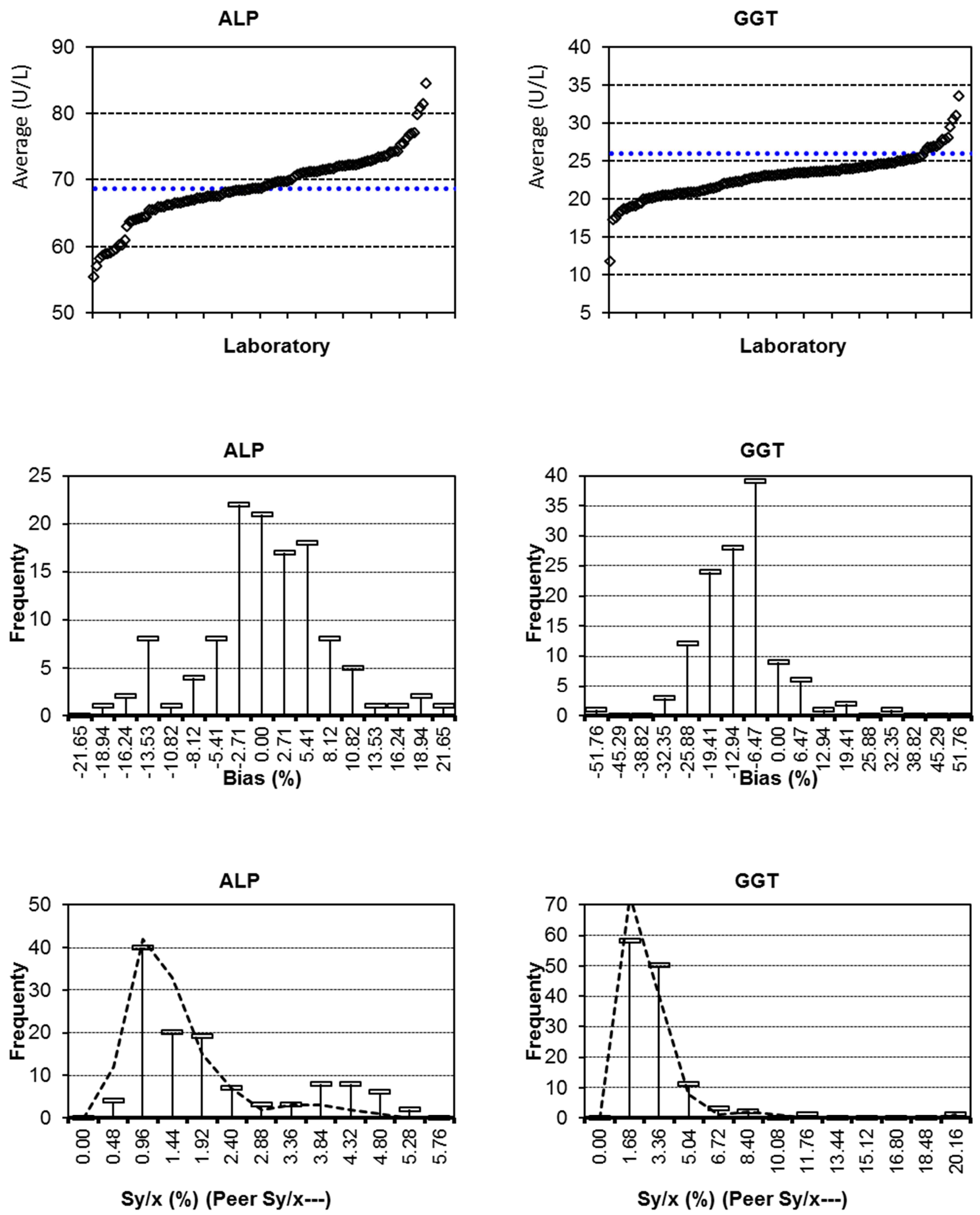
Supplemental Figure 2 (part 1 – ALT with and without PP). Distribution of laboratory averages (blue dotted line represents the RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.



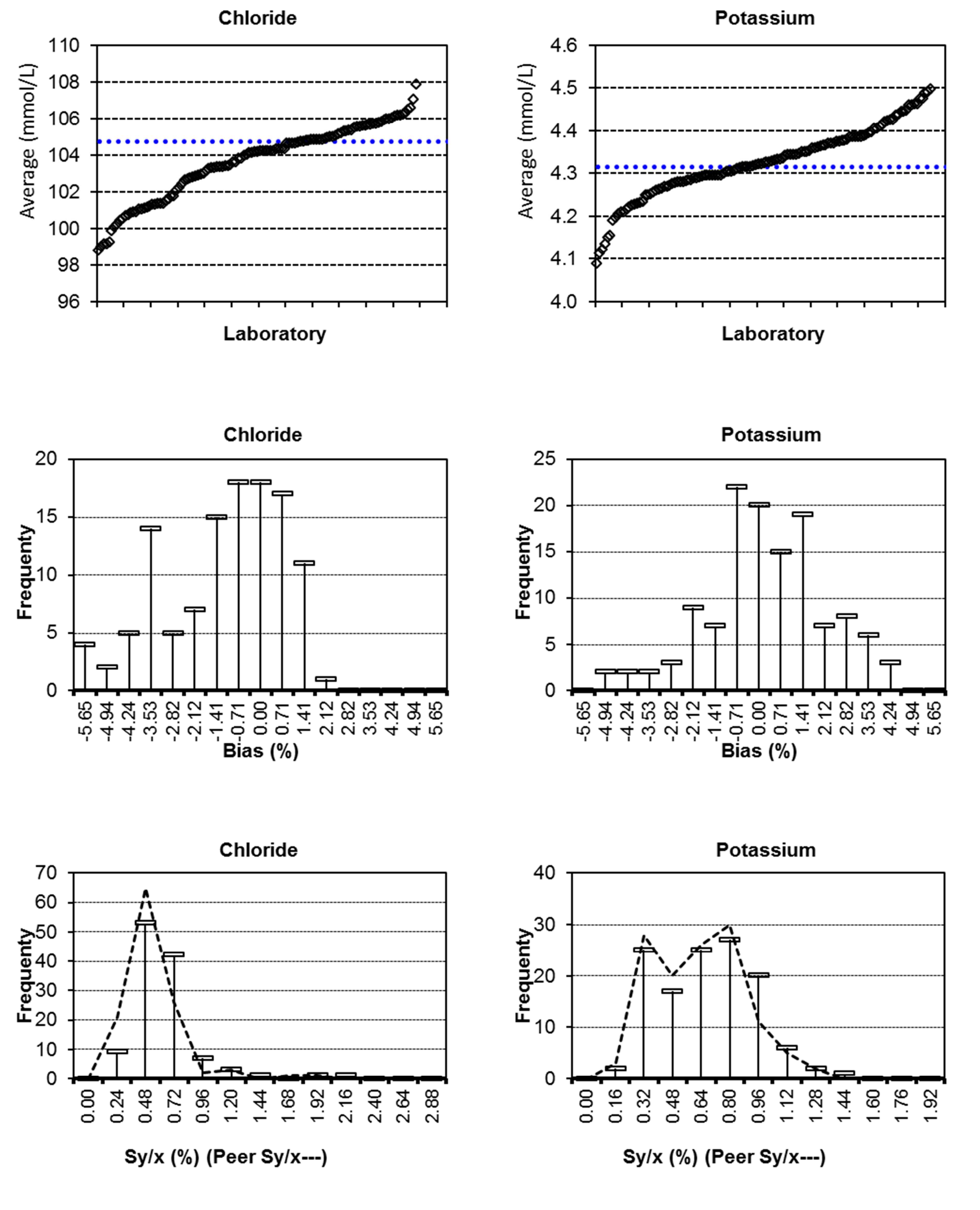
Supplemental Figure 2 (part 2 – AST with and without PP). Distribution of laboratory averages (blue dotted line represents the AMTM/RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.



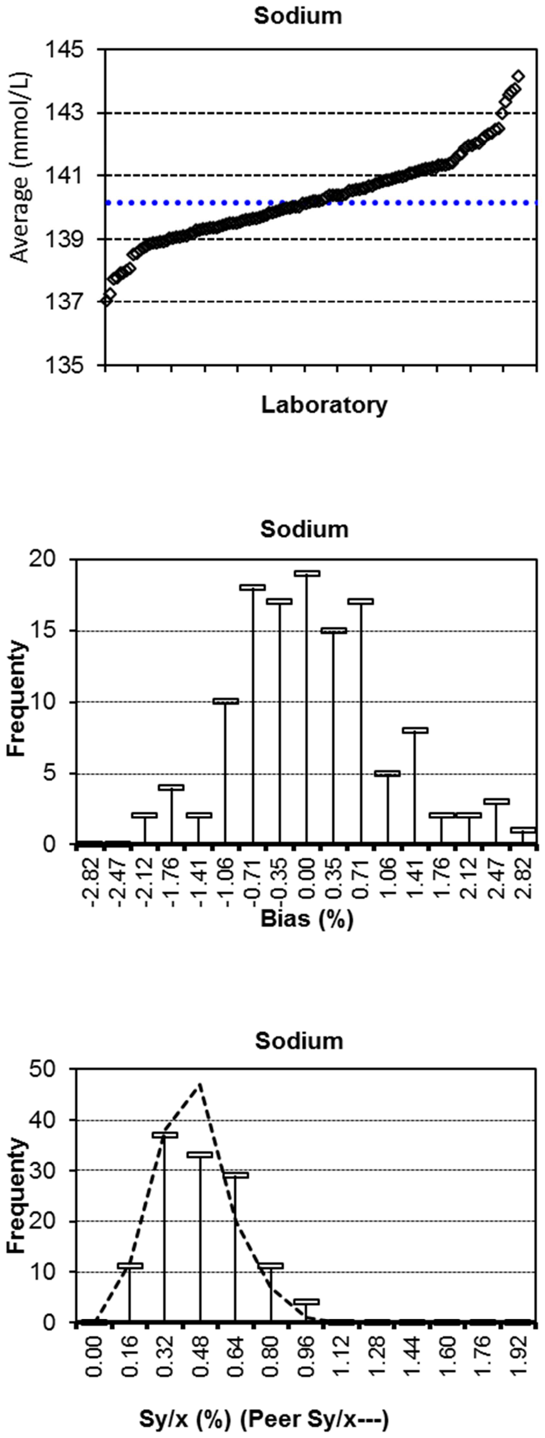
Supplemental Figure 2 (part 3 – LDH forward (FW) and reverse (RV)). Distribution of laboratory averages (blue dotted line represents the AMTM/RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.



Supplemental Figure 2 (part 4 – ALP and GGT). Distribution of laboratory averages (blue dotted line represents the AMTM/RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.

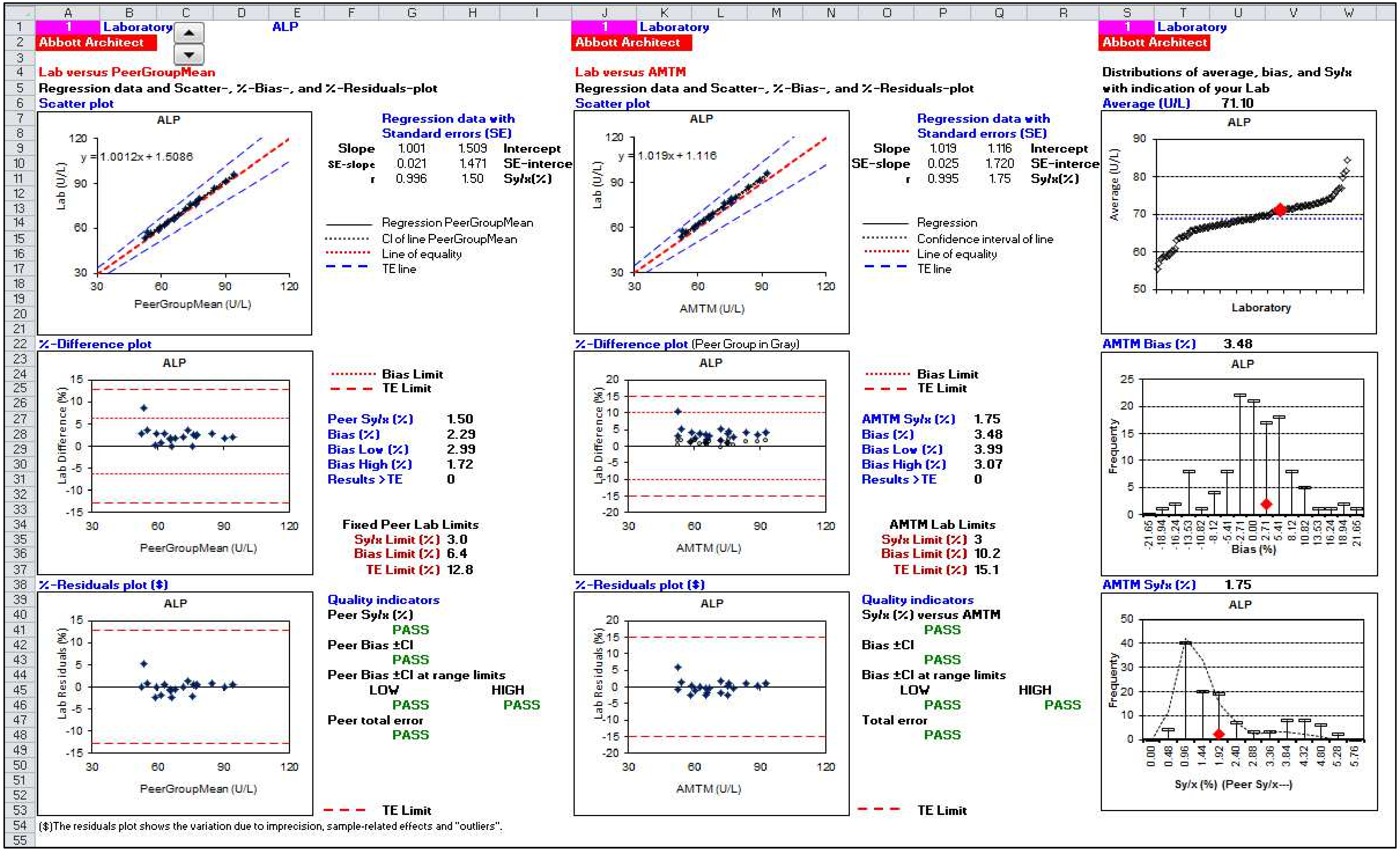


Supplemental Figure 2 (part 5 – chloride and potassium). Distribution of laboratory averages (blue dotted line represents the AMTM/RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.



Supplemental Figure 2 (part 6 – sodium). Distribution of laboratory averages (blue dotted line represents the AMTM/RMP-corrected AMTM), AMTM/RMP-corrected AMTM bias and AMTM/RMP-corrected AMTM Sy/x.

Screenshot of the analyte-specific Excel files (page 1-3):



1 Laboratory ALP

Abbott Architect

Lab versus PeerGroupMean

Regression data and Scatter-, %-Bias-, and %-Residuals-plot

Scatter plot

Regression data with Standard errors (SE)

Slope	1.001	1.509	Intercept
SE-slope	0.021	1.471	SE-interce
r	0.996	1.50	Sy/x(%)

Regression PeerGroupMean
CI of line PeerGroupMean
Line of equality
TE line

%-Difference plot

Bias Limit
TE Limit

Peer Sy/x (%)	1.50
Bias (%)	2.29
Bias Low (%)	2.99
Bias High (%)	1.72
Results >TE	0

Fixed Peer Lab Limits

Sy/x Limit (%)	3.0
Bias Limit (%)	6.4
TE Limit (%)	12.8

%-Residuals plot ($)

Quality indicators

Peer Sy/x (%)
PASS

Peer Bias ±CI
PASS

Peer Bias ±CI at range limits
LOW HIGH
PASS PASS

Peer total error
PASS

TE Limit

($)The residuals plot shows the variation due to imprecision, sample-related effects and "outliers".

1 Laboratory

Abbott Architect

Lab versus AMTM

Regression data and Scatter-, %-Bias-, and %-Residuals-plot

Scatter plot

Regression data with Standard errors (SE)

Slope	1.019	1.116	Intercept
SE-slope	0.025	1.720	SE-interce
r	0.995	1.75	Sy/x(%)

Regression
Confidence interval of line
Line of equality
TE line

%-Difference plot (Peer Group in Gray)

Bias Limit
TE Limit

AMTM Sy/x (%)	1.75
Bias (%)	3.48
Bias Low (%)	3.99
Bias High (%)	3.07
Results >TE	0

AMTM Lab Limits

Sy/x Limit (%)	3
Bias Limit (%)	10.2
TE Limit (%)	15.1

%-Residuals plot ($)

Quality indicators

Sy/x (%) versus AMTM
PASS

Bias ±CI
PASS

Bias ±CI at range limits
LOW HIGH
PASS PASS

Total error
PASS

TE Limit

1 Laboratory

Abbott Architect

Distributions of average, bias, and Sy/x with indication of your Lab

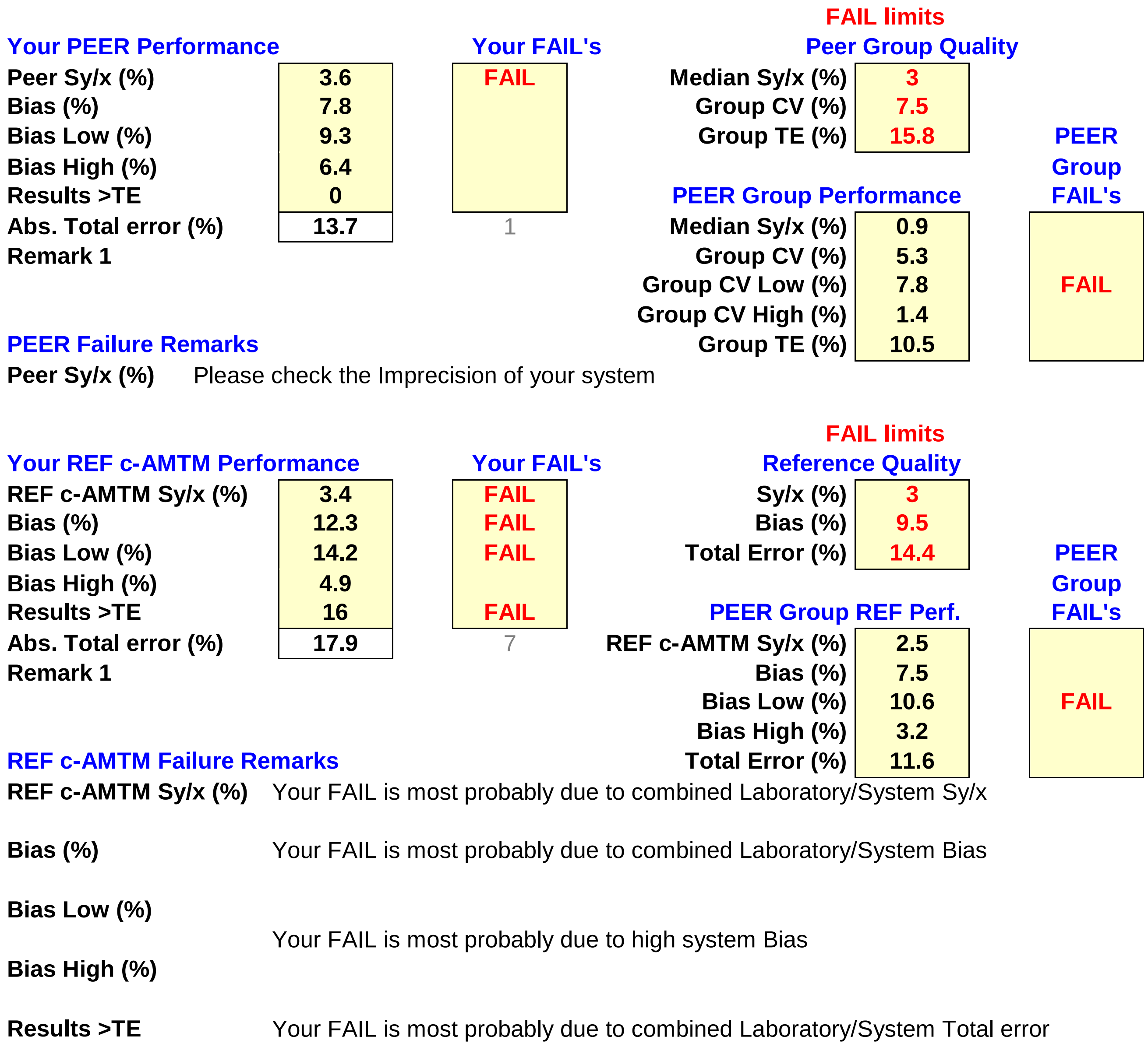
Average (U/L) 71.10

AMTM Bias (%) 3.48

AMTM Sy/x (%) 1.75

Extract from the laboratory guide for data interpretation - example of page 4 on the analyte-specific Excel files:

PASS/FAIL Example 1



Your PEER Performance		Your FAIL's
Peer Sy/x (%)	3.6	FAIL
Bias (%)	7.8	
Bias Low (%)	9.3	
Bias High (%)	6.4	
Results >TE	0	
Abs. Total error (%)	13.7	1
Remark 1		

FAIL limits

Peer Group Quality	
Median Sy/x (%)	3
Group CV (%)	7.5
Group TE (%)	15.8

PEER Group Performance		PEER Group FAIL's
Median Sy/x (%)	0.9	
Group CV (%)	5.3	
Group CV Low (%)	7.8	FAIL
Group CV High (%)	1.4	
Group TE (%)	10.5	

PEER Failure Remarks

Peer Sy/x (%)	Please check the Imprecision of your system

Your REF c-AMTM Performance		Your FAIL's
REF c-AMTM Sy/x (%)	3.4	FAIL
Bias (%)	12.3	FAIL
Bias Low (%)	14.2	FAIL
Bias High (%)	4.9	
Results >TE	16	FAIL
Abs. Total error (%)	17.9	7
Remark 1		

FAIL limits

Reference Quality	
Sy/x (%)	3
Bias (%)	9.5
Total Error (%)	14.4

PEER Group REF Perf.		PEER Group FAIL's
REF c-AMTM Sy/x (%)	2.5	
Bias (%)	7.5	
Bias Low (%)	10.6	FAIL
Bias High (%)	3.2	
Total Error (%)	11.6	

REF c-AMTM Failure Remarks

REF c-AMTM Sy/x (%)	Your FAIL is most probably due to combined Laboratory/System Sy/x
Bias (%)	Your FAIL is most probably due to combined Laboratory/System Bias
Bias Low (%) Bias High (%)	Your FAIL is most probably due to high system Bias
Results >TE	Your FAIL is most probably due to combined Laboratory/System Total error

Peer FAILS

You FAIL the Peer Sy/x, which indicates precision problems because most systems operate with Peer Sy/x values <3%.

The Peer Group itself has a high variation at the low concentration (see Group CV Low (%)); this puts you at risk to FAIL the comparison with the reference (indeed, you also FAIL the REF low bias limit).

REF c-AMTM FAILS

You FAIL most of the REF performance limits. Some of them are most probably a combination of your own and your systems performance. The “Low Bias FAIL”, however, is most probably due to system bias at low concentration; also the systems shows a FAIL there.

5 Chapter V

The Empower Project

This chapter is based upon:

- De Grande L, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont L. The Empower Project – A new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53:1197-204.
- Goossens K, Stöckl D, Thienpont L. Empower IVD ● Globe - a status update (meeting report). http://www.westgard.com/empower-ivd-globe-2014.htm

Highlights

Each the Percentiler, Flagger and Master Comparison are useful tools for the assessment of IVD test stability and comparability. In order to inform stakeholders – both laboratories and manufacturers – on the different aspects and outcomes of these studies, a communication platform was created which is called the Empower Project.

Highlights of the research:

- The Empower project is the overarching communication platform on different tools with utility to assess the quality of IVD assays.
- The Empower project informs interested parties worldwide on its progress and perspectives by the means of oral and written communication (conferences, posters, industry-sponsored workshops, etc.).
- The Empower project informs participating parties on the outcome of the different tools it covers, by the means of reports, both individual and general, and peer-reviewed publications.

1. Introduction – What is the Empower project?

The primary objective of the Empower project is not only to assess, but also to improve the comparability and stability of laboratory tests. To accomplish these goals the project was founded on four pillars (see Figure 1):

1. The Master Comparisons (see chapter 4).
2. Virtual EQA 1: mid- to long-term monitoring of patient percentiles using the Percentiler (see chapter 2) and the Flagger (see chapter 3).
3. Virtual EQA 2: mid- to long-term monitoring based on IQC data. This pillar was not further investigated.
4. Conceptual and statistical education tailored to the analytical quality needs of the clinical laboratory.

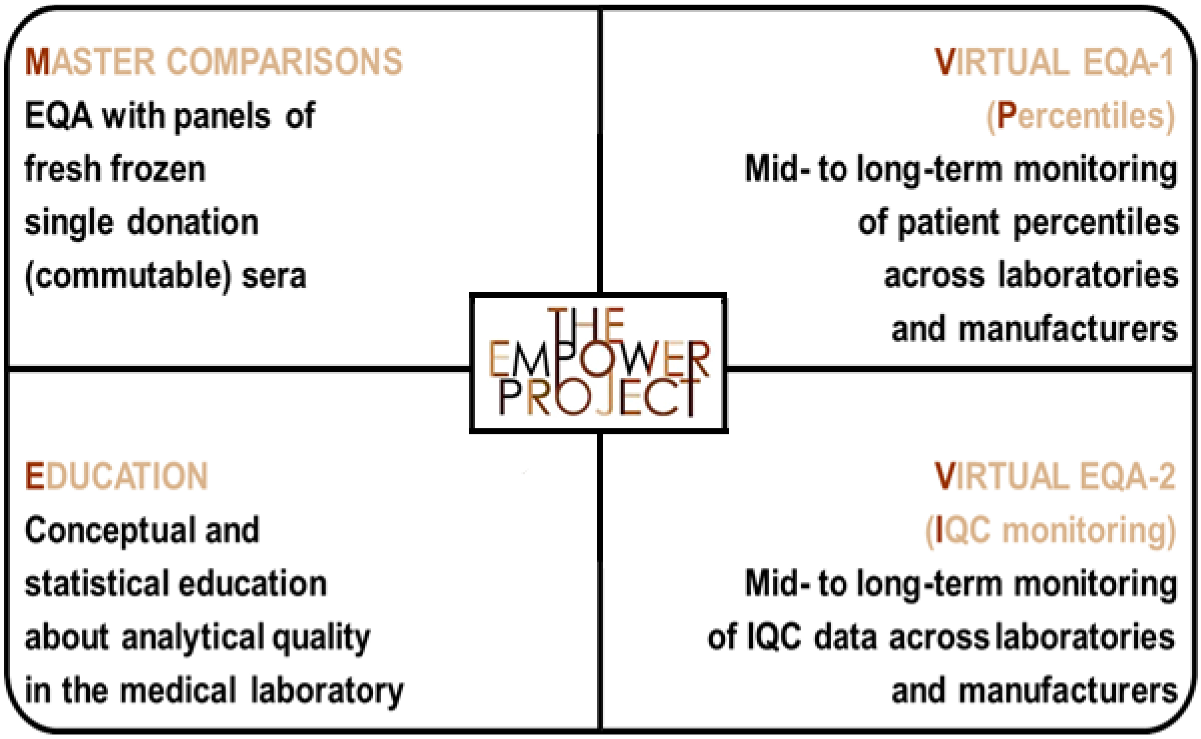


Figure 1. The four pillars of the Empower project.

To manage the goal of quality improvement, the Empower project aims at establishing a bottom-up cooperation between stakeholders interested in quality of performance, i.e. laboratories and IVD manufacturers. The project organizers, thereby, try to act as mediator between all involved parties. Through this cooperation, reliable evidence should be created about quality and stability of laboratory tests from data which are appealing enough to be translated into action by the above stakeholders. The work which has been conducted in the framework of the Empower project so far, is summarized in the publication on pages 207-218.

2. Communication with the participants

Role of communication in the Empower project

A key aspect of the Empower project is sustained contact with all stakeholders (manufactures, laboratories, LIS providers). This communication is vital to keep the interested parties informed on the major findings of the project and is important for the recruitment of new participants.

Forms of communication

To make laboratories aware of the goals and work of the Empower project, both written and oral communication was used:

- Industry sponsored workshops were held at the IFCC Worldlab – 22nd international congress of clinical chemistry and laboratory medicine (22-26 June 2014) in Istanbul (Turkey) and at EuroLabFocus – 3rd EFLM-UEMS congress (7-10 October 2014) in Liverpool (UK).
- Posters on different aspects of the Empower project were submitted and presented at congresses worldwide.
- The project was presented on GLIMS user meetings in Paris, Utrecht and Ghent.
- Manuscripts and invitations were published in different journals. “The Bottom Up Approach to Quality Assurance” was published in The Pathologist (1) and was translated by Dr. Hassan Bayat in Farsi for an Iranian magazine called the Laboratory News magazine in a special issue for the Tehran Quality Improvement Congress (extract available in Annex on page 220). A Czech magazine, called the FONS bulletin, published an invitation to participate to the Percentiler (2; extract available in Annex on page 219).
- Flyers were distributed at every possible occasion. An extract of this flyer can be found in the Annex (page 222).

In addition, stakeholders were kept informed on the progress and findings of the project:

- Participants were contacted individually with both short communications and more extensive reports which addressed the quality of their own IVD systems compared to their peer group and to the state-of-the-art quality.
- Global reports were written which addressed several aspects of the Empower project going from the project status of both the Percentiler and Flagger, and explanation on data analysis, to a comparison of the Percentiler findings with the Master Comparison results, and a summary on quality-related issues for the different peer group (peer

group reports). All these reports were made available on www.stt-consulting.com (Empower tab).

- User meetings were held in Ghent on December 10th 2014 and on December 9th 2015. The aim of these meetings was to report on the status of the Empower project. As mostly users of the project had registered to attend the meeting, it was the intention to have a lively discussion: to learn what they think of the project, and how they use it; to know what they appreciate in the project, what is missing or what they would like to see changed. The meeting report of the first meeting was published as a guest assay on the Westgard website (3; extract available in Annex on page 221).

Last but not least, several publications in peer-reviewed journals were published:

- Manuscript in Clin Chem Lab Med: “The Empower project – a new way of assessing and monitoring test comparability” (4).
- Editorial in Clin Chem Lab Med: “A new integrated tool for assessing and monitoring test comparability and stability” (5).
- Percentiler manuscript in Clin Chim Acta: “Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make “Big Data” work” (6).
- Letter to Editor in Clin Chem Lab Med: “On-line flagging monitoring – a new quality management tool for the analytical phase” (7).
- Manuscript in The Pathologist on the use of the Percentiler and Flagger for FT4 and TSH: “Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies” (8).

References

1. Thienpont L, Stöckl D. The Bottom-Up Approach to Quality Assurance. The Pathologist 2014;314:28-30.
2. Goossens K, Thienpont L, Stöckl D, Van Uytfanghe K. The Percentiler (Patient Percentile Monitoring) – Invitation to participate. Bulletin FONS;1:9-12
3. Goossens K, on behalf of the Empower team. Guest Essay - Empower IVD Globe Meeting Report. https://www.westgard.com/empower-ivd-globe-2014.htm (accessed January 2015).
4. De Grande L, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont L. The Empower Project – A new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53(8):1197-204.
5. Secchiero S, Plebani M. A new integrated tool for assessing and monitoring test comparability and stability. Clin Chem Lab Med. 2015 Jul;53(8):1131-3.
6. Goossens K, Van Uytfanghe K, Twomey P, Thienpont L, and Participating laboratories. Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make “Big Data” work. Clin Chim Acta 2015;445:12-18.
7. Goossens K, Brinkmann T, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Clin Chem Lab Med 2015;53(10):e269-70.
8. De Grande LAC, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru M, Thienpont LM, for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta 2016; doi: 10.1016/j.cca.2016.04.032. [Epub ahead of print].

Article 6. The Empower Project

The Empower Project – A new way of assessing and monitoring test comparability and stability

Linde AC De Grande[1], Kenneth Goossens[1], Katleen Van Uytfanghe[1], Dietmar Stöckl[2], Linda M Thienpont[1*]

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
[2]STT-Consulting, Horebeke, Belgium

Introduction

Manufacturers and laboratories have common interest in precise, unbiased, and stable in vitro diagnostic assays enabling optimal patient care. Although they both monitor the above test attributes, they have different objectives and access to existing data, which are facts that might hamper the dialogue between them. For example, manufacturers are mainly interested in the global performance of their assays (= peer performance), while laboratories rather focus on their own performance. However, for trouble-shooting purposes, peer performance is also of interest to laboratories. Manufacturers monitor laboratories by an online link with their systems, while laboratories have easy access to their own data. The data sources can be bridged by independent third party programs for peer group based combined internal quality control (IQC)/external quality assessment (EQA). However, this approach has limitations. Commutability issues of the used materials make that peer group assessment cannot give information on trueness of performance. Additionally, it may cause that variations in patient data (e.g., trends and shifts due to reagent lot changes) are not well reflected (1-3). Besides, continuous monitoring of the results is rather the exception and, even if done, the data are usually not accessible in real-time. In addition, the external program providers mostly do not critically review or publish the data, but leave the interpretation to the participating laboratories. This practice is of course driven by the commercial surrounding in which they operate, which hampers them to disclose performance data of individual diagnostic manufacturers. In contrast, independent national or regional EQA schemes theoretically are in the position to openly demonstrate the performance of commercial test systems. However, this requires that sufficient laboratories participate, so that the peer groups can be well defined. This is for most of the schemes not possible, therefore, they rather restrict to assessing the competence of laboratories (2, 3). This is in turn limited by the fact that EQA schemes seldom work with fully commutable materials, conduct surveys at low frequency and report retrospectively. From this perspective, it would be desirable to implement an independently operated "online" tool that enables to monitor comparability and stability between peer groups and laboratories without being confounded by non-commutability issues, because it uses real patient samples. To maximize the utility of the tool, the information should be shared between participants and manufacturers but within confidentiality constraints. This means that an individual evaluation report should only be available to the laboratory to which it applies. The tool could in the same time serve to empower laboratories for the future tasks they face, among others, providing input for the development and implementation of global health-care policies.

In response to these needs, we initiated the so-called "Empower" project. It is intended to establish a bottom-up cooperation between laboratories and manufacturers, so that they can pursue the common objective of assessing and improving test comparability and stability, whereby we see our role as independent third party mediator. It is our strong belief that such a transparent cooperation will be of benefit to all stakeholders involved in laboratory medicine. The project stands on four pillars: i) master comparisons with fresh-frozen single-donation serum samples; ii) monitoring of patient percentiles and iii) IQC, both across laboratories and manufacturers; iv) conceptual and statistical education about analytical quality in the medical laboratory (e.g., analytical performance specifications) and elaboration of statistically sound and "actionable" experiments for analytical quality management and assurance. Laboratories are free to participate in all pillars of our project, or to select the most appropriate one(s) for their purpose. Here, we report on the status of the project with respect to the master comparisons and patient percentile monitoring, and share the first observations on test comparability and stability of performance.

Materials and methods

Master comparisons

As previously described, we conduct the master comparison surveys for diagnostic assays with panels of 20 fresh frozen, single-donation, commutable serum samples (each available in a volume of ~180 mL) (4-6). The samples are prepared by the Clinical and Laboratory Standards Institute C37-A protocol, however, without pooling and filtration, and are dispatched on dry ice (7). Participation is made conditional of the use of a homogeneous test system, i.e., instrument, reagent and calibrator from the same manufacturer.

The number and selection of laboratories is adapted to obtain peer groups representing the main manufacturers/ diagnostic test systems (~20 laboratories per manufacturer/ system). Participation also includes the in-house laboratories of the respective manufacturers. For each survey, we select 8 different analytes from the clinical chemistry test menu of modern platforms (for the analytes covered up to now, see Table 1). The participants are requested to do the measurement in singlet under within-run conditions. The quality of assays and laboratories is assessed from 4 quality indicators at the peer group and "reference" level. The latter uses either the all manufacturer trimmed mean (AMTM) or reference method values as target: i) the standard error of the estimate (% Sy/x) from linear regression analysis; if data are compared to the peer group mean, the Sy/x is a measure for within-run imprecision, if compared to the reference target, it reflects the combined imprecision (both random and sample related effects); ii) bias (%) at the mean concentration and the range limits (low and high concentration end); iii) total error (%); iv) the number of results observed outside the total error limits. These estimates are tested against a hierarchy of decision limits, i.e., limits that account for state-of-the-art performance, but, also limits derived from biological variation data (8).

Table 1. Analytes covered in the master comparisons (MC) and patient percentile monitoring (PPM)

Alanine aminotransferase[a]	Glucose[c]
Albumin[b]	Lactate dehydrogenase[a]
Alkaline phosphatase[a]	Magnesium[b]
Aspartate aminotransferase[a]	Phosphate[c]
Calcium[b]	Potassium[a]
Chloride[a]	Sodium[a]
Total cholesterol[c]	Total protein[b]
Creatinine[c]	Total triglycerides[c]
Gamma-glutamyl transferase[a]	Uric acid (urate)[c]
Analytes only covered in the MC	**Analytes only covered in the PPM**
HDL-cholesterol[c]	C-reactive protein
LDL-cholesterol[c]	Total-bilirubin

a, b, c Analytes covered in references (10), (5) and (6), respectively.

Patient percentile monitoring

We monitor the daily medians of the results for 20 commonly measured analytes in serum or plasma. All types and sizes of laboratories can participate. The laboratories calculate instrument-specific daily medians from outpatient results and send the data by e-mail to our database. Several vendors of laboratory information systems offer cost-free solutions for automatic calculation and electronic transfer. Alternative solutions are extraction of weekly/monthly data from the system and shipment in batch. Formats readable in our database are an e-mail embedded table, Excel-files, and text-files . Note that we do the mapping of the laboratories' mnemonics for the different analytes and units for expression of the medians. Via a user interface with authentication (access by user name and password) for secured authorization, the participating laboratory can plot for each analyte the course of the moving median. If a laboratory reports medians for different instruments, the moving medians (instrument-specific colored) are shown in the same plot. For interpretation, preliminary desirable limits for mid- to long-term bias are included. These are guided by biological variation and state-of-the-art performance. The user application allows selection of i) the number of consecutive medians (n = 5, 8, 16) used for calculation of the moving median, ii) time window and iii) exclusion of data from weekends. Each plot also shows the long-term median of the concerned individual laboratory, as well as the peer group or all devices median (freely to select). Additional numerical information is provided on the long-term imprecision (the so-called "robust CV", %), and the bias calculated in comparison to the peer group or all devices target, as well as a "desirable" target. Currently we use the medians of the reference intervals determined in the trueness-based "Nordic Reference Interval Project (NORIP)" as preliminary reference source for that target (9). The user can download and print the plots. He has also access to his own entries in the database with the possibility to filter/sort according to analyte/date. This facilitates tracing back on which date graphical aberrant observations started. The graphical user interface can be accessed at https://www.thepercentiler.be/ (to see the demo version, log in with "demolab" as username and "demo1234" as password).

Results

Status of the project

Results of the master comparison surveys conducted up to now are described elsewhere (4-6, 10). In the most recent survey (2014) a total of 125 laboratories from 21 different countries (15 in Europe, and Australia, Canada, Malaysia, South-Korea, Singapore and the USA) participated. The 5 main manufacturers also joined with their in-house laboratories (Abbott, Beckman, Ortho, Roche and Siemens). In the patient percentile monitoring part, currently 100 laboratories from 15 different countries (11 in Europe, and Australia, India, Russia, the USA) are participating with a total of 182 devices. Most of the test systems involved in the 2014 master comparison survey are also represented in percentile monitoring. Table 1 shows that most analytes covered in the master comparisons (20 until now) are also addressed in patient percentile monitoring.

Test performance, comparability across manufacturers and laboratory performance

As described elsewhere in detail, the design of the master comparisons with 20 single-donation commutable samples allows to assess different performance attributes of the examined assays, and also individual laboratory performance (4-6). Apart from some exceptions, assay peer group assessment showed a good intrinsic analytical quality in terms of within-run and combined imprecision, and total error. It also demonstrated sufficient robustness for satisfactory performance in a daily laboratory context. However, there was room for improvement at higher and lower concentrations. Assessment at the reference level showed for several analytes good comparability between manufacturers/assays, e.g., for total protein, cholesterol, glucose, phosphate and uric acid (5, 6), while for others considerable calibration differences were obvious, e.g., for albumin (5). Particularly striking in this regard were the biases against the targets for enzymes set by the IFCC reference methods (10-15). Also long-term assay drift/uncorrected biases for a single manufacturer were sometimes uncovered, e.g., magnesium, creatinine, LDL-cholesterol, phosphate, uric acid and chloride in (5, 6, 10). Assessment against the reference method or AMTM showed for most assays and analytes sufficient analytical specificity, but for others vulnerability to sample-related effects, e.g., HDL- and LDL-cholesterol in (6). The bias limits used for assessment demonstrated that for certain analytes the state-of-the-art is such that most assays, apart from some, can meet the desirable biological variation bias specifications (e.g. for total protein, phosphate, triglycerides, uric acid, alkaline phosphatase, and potassium (5, 6, 10)). For some biologically more tightly regulated analytes, the biologically inferred limits are not feasible, e.g., glucose, cholesterol and chloride (6), or would require improvement of lot-to-lot consistency, e.g., calcium (5). In contrast, sodium assays showed exceptionally well performing, almost within the tight biological bias limit (10). Assessment of the laboratory performance strikingly showed that sometimes large between-laboratory differences (>30%) occurred for all analytes (6, 10).

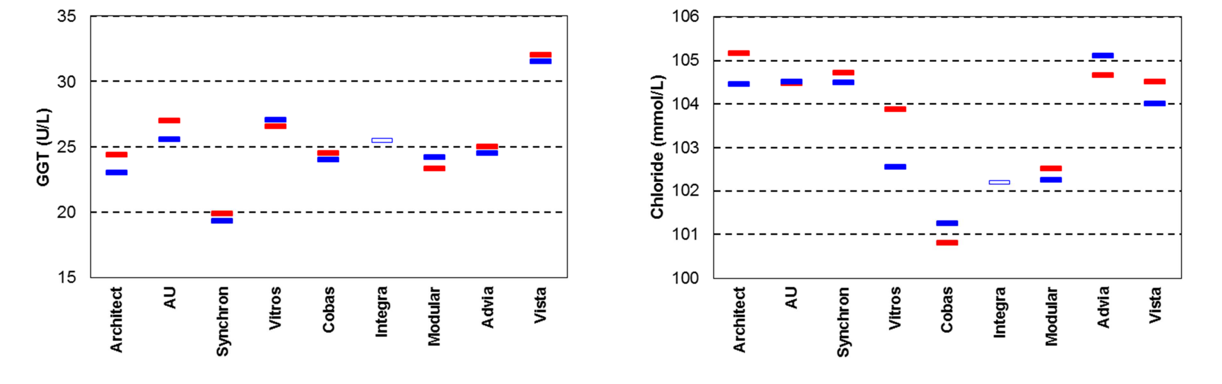


Figure 1. Illustration of the match between the peer group means (red rectangles) in the 2014 survey of the master comparisons and the median values (blue rectangles) in patient percentile monitoring for gamma-glutamyl transferase (GGT) and chloride.

These discrepancies could partly be ascribed to the biases in the used assays, but likely also point to severe laboratory effects on performance of assays in daily practice.

Similar observations were made from the patient percentile monitoring data. For example, the median values matched the aforementioned calibration differences revealed for gamma-glutamyl transferase and chloride in the 2014 master comparison survey (Figure 1) (10). Indeed the gamma-glutamyl transferase moving median values ranged from ~20 to ~32 U/L, those for chloride from ~101 to ~105 mmol/L.

Stability of laboratory/test performance

First results from patient percentile monitoring show that laboratories with high daily throughput and/or low variation in patient population typically perform with low variation and mostly good concordance between the different instruments. Other laboratories have a higher long-term variation in performance. If this is due to a lower throughput or higher population variation (typical for laboratories operating in a medium-size hospital), the variation can partly be reduced by selecting a higher n for calculation of the moving median. Other observations are about drifts or shifts, or transient to long-term bias, e.g., between different instruments used in a laboratory, of one particular instrument compared to the others, or of the laboratory compared to its peer. Interestingly, shifts or drifts sometimes apply for several laboratories belonging to the same peer, which confirms that they are caused by a major manufacturer event, e.g., a reagent or calibrator lot change (Figure 2A and B). In other cases, laboratories can relate the observed instability to a calibration event (example shown in Figure 2C).

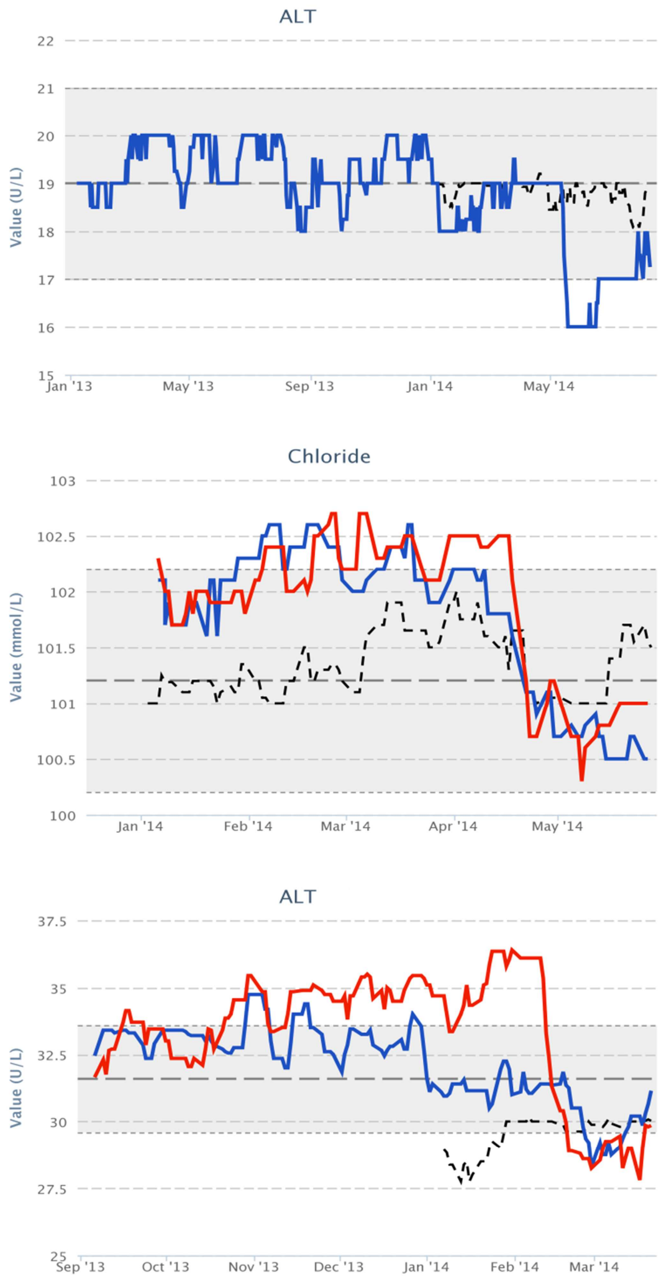


Figure 2. Significant test instability for alanine aminotransferase (ALT) and chloride due to a confirmed reagent lot change (A), a calibrator lot change (B) and a laboratory calibration event (C). In (A) and (B) it is illustrated how lot changes can disturb the stable performance. The long-broken grey line represents the median calculated from all daily medians provided by the laboratory to which the graph applies. In (C) the moving median for one of the instruments (red colored full line) started to drift around the 20th of December 2013, and on February the 12th 2014, both instruments (also the blue one) were recalibrated by the laboratory, which caused in both a shift. The shifts moved the medians outside the stability zone (shaded area between short-broken grey lines). The black short-broken line represents the peer group moving medians in (A), (B) and (C).

Although certain observations can rather easily be explained, longer observation times and more solid peer groups are needed for a systematic investigation of the root causes.

Discussion

The Empower project is an integrated quality assurance tool for laboratories and manufacturers. Its unique design based on real patient results allows to assess/demonstrate quality aspects without being confounded by commutability issues (16, 17). It facilitates remediating actions, because it reveals major bias components/sources, such as the manufacturer (assay), laboratory, instrument, the reagent/calibrator lot and recalibration by the laboratory itself (Figure 3).

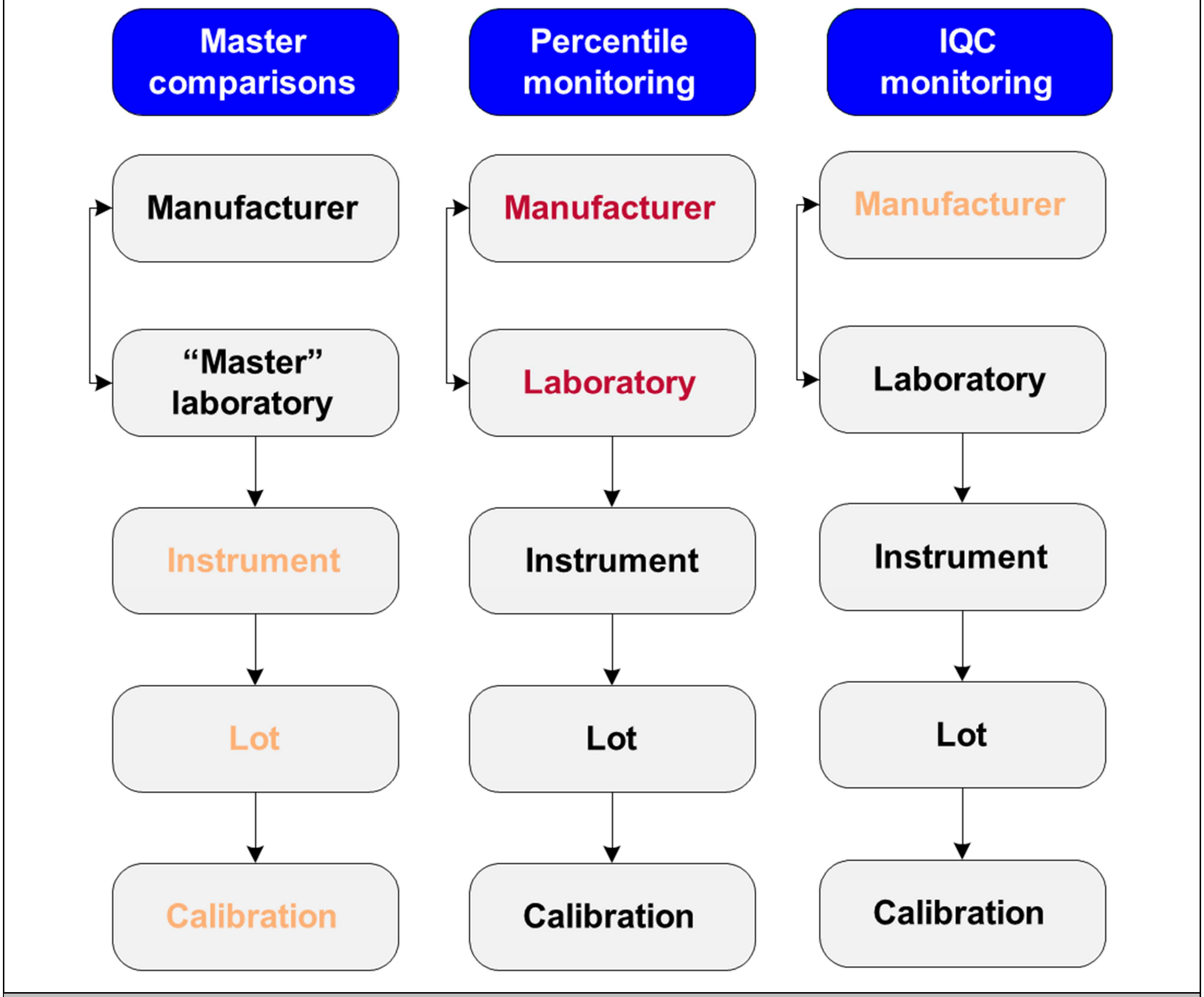


Figure 3. Assessment (and control) of bias components/sources. Components in black can be assessed by the above standing pillar, those in orange cannot; those in red probably also can in high-throughput laboratories that mainly work with general practitioners (samples almost exclusively from outpatients).

The focus of the master comparisons, which are conducted across assays and laboratories, is on how well the intrinsic analytical quality of assays on release by the respective manufacturers is reproduced by the end users under “field” conditions all over the world. Target setting is based on reference method measurements or the AMTM. These targets allow to assess either the real traceability (standardization status) or the comparability between assays and laboratories. The information on traceability is of utmost use for the discipline of clinical chemistry to investigate the extent of implementation of standardization efforts. Note in this regard the striking example of the bias observed for enzyme assays. For the individual laboratory it is first-hand information that can help in decisions on the acquisition of new instruments. As such, the master comparisons provide the participating laboratories with a reliable calibration fix-point of their own performance within the peer group and of the latter compared to other peers. Naturally, this is only a point-estimate in time that should continuously be monitored. This is where patient percentile (and IQC) monitoring comes into play. Indeed, the stability of the peer group calibration fix-points can be appreciated from concordant medians from outpatient results (Figure 1). In addition, laboratories can use their medians as a tool to monitor the mid- to long-term stability of their own calibration status, again in comparison to their peer, and/or to uncover shifts/drifts and the sources thereof (18). Of course this requires that the moving medians in time truly reflect the analytical variation, without being confounded by other sources of variation. In a pilot study we showed that by working with medians from outpatients and omitting medians from weekends and holidays (days with lower throughput and/or altered ratios of in- to outpatients), the effect of patient population variability can be suppressed. We inferred this from a congruent course in time of the moving medians and mean of daily IQC data (18). Meanwhile it is our experience that in high throughput laboratories mainly serving outpatients, the moving medians can be calculated from a low number of daily medians (n = 5). This is the ideal number for detection of analytical instabilities (shifts, drifts). In contrast, for laboratories in a hospital context, a higher n is required to partly compensate for the effect of a more variable patient population and lower throughput. We offer in the user interface n = 8 or 16, however, the latter is the limit to prevent too much smoothing and loss of resolution. Another asset of the percentile monitoring design is that it shows the instrument-specific stability in one plot. This allows the laboratories to monitor the interchangeability of results among different instruments, and/or detect the occurrence of instrument-specific special events.

Notwithstanding the above potential of the percentile monitoring tool, we recommend the users to do the interpretation with caution. Indeed, certain influential factors may explain aberrant or more variable medians. We learned, for example, that in hospital laboratories dialysis or oncology patients are often registered as outpatient, and that their samples are preferentially measured on one instrument. Note, however, that by closely working with our participants, we can share our experience to enable more critical interpretation. We also recommend sample exchange between partner-laboratories belonging to the same peer group and, preferably, participating in patient percentile monitoring, since this may be very helpful to exclude or confirm observed laboratory biases.

We want to emphasize that monitoring of patient medians is not a substitute for daily IQC. We advocate it as a complementary observation tool from patient data that can cover much longer observation times.

A fundamental question in all parts of the Empower project is whether the observed differences in quality of performance or instability are to be considered significant. This points to the importance of performance specifications for meaningful conclusions (19-22). In the absence of a consensus on this topic, we use preliminary limits that are guided by biological variation (8), and also by state-of-the-art performance. This means that for tightly regulated analytes we expand the limits based on biological variation to account for the current quality offered by manufacturers. Note that for patient percentile monitoring we express the limits for allowable bias in absolute terms (tailored to the used SI-units). The reason is that this allows us to show them in the user interface as so-called stability limits that should not be exceeded by longer than 1 week. See, for example, the shaded zone between 17 and 19 U/L (median ± 2 U/L) in Figure 2A for ALT. Interestingly, we found the patient percentile monitoring an excellent tool to test how realistic our quality goals are, e.g., the stability limit of 1 mmol/L for sodium (23). For other analytes with very high biological variation, such as C-reactive protein, we set a general upper limit of ~10%.

Another important question is which targets to use. For the master comparisons part, this is discussed elsewhere (6). In the percentile monitoring part we compare the medians in first instance with the peer group medians, but also with a reference median. We use the median from the NORIP reference intervals, which is, to the best of our knowledge, the only source that claims to be “trueness-based” (9). The reliability is high for analytes such as sodium and calcium, but the information for some enzymes has to be interpreted critically. There have been changes in the IFCC recommended methods and it is known that these are either not carefully or uniformly adopted by manufacturers. Therefore, we still consider NORIP as a preliminary reference source and will follow up, e.g., by cross comparison with the reference interval information from manufacturers, and new projects.

Of course, the utility of our project has to be improved on a continuous basis. For example, we aim at a platform that stimulates the dialogue on a basis of trust between the participant laboratories and manufacturers. We work on this by establishing close contacts with both parties. We also plan to develop a new tool that investigates the effect of analytical (in)stability on a surrogate medical outcome, such as the frequency of “flagged results” (18). Together with realistic quality goals that result in meaningful conclusions, this tool might be an excellent basis to strengthen the physician/laboratory interface by more transparent communication on performance. The Empower database potentially can become a source for “big data mining” with utility for studies that relate the outcome of therapeutic strategies to median values in patient cohorts (e.g. the Dialysis Outcomes and Practice Patterns Study) (24). From the perspective that the project’s general emphasis is on interchangeability of laboratory results, it can potentially also contribute to modern clinical needs such as the definition of common reference intervals or clinical decision limits, implementation of electronic health records, and development of evidence-based clinical practice guidelines for application of consistent standards of medical care.

Conclusions

The Empower project provides evidence on the intrinsic quality of assays and how this quality is sustained under field conditions. It also demonstrates how well assays and laboratories compare, and how stable they perform. In addition, it enables to uncover all major bias components/sources. The major asset of the project is that it works with data generated from real patient samples, and can be linked to observations in daily IQC practice. From this perspective, we believe it is a new integrated tool for modern quality management of benefit to all stakeholders with interest in reliable laboratory data. It can help the discipline of clinical chemistry to derive realistic quality specifications, and can strengthen the laboratory/manufacturer dialogue and laboratory/physician interface. Ultimately, if the evidence provided by the project is translated into action by laboratories and manufacturers, it can contribute to a yet to be established translational laboratory medicine and better patient care.

References

1. Miller WG. The role of proficiency testing in achieving standardization and harmonization between laboratories. Clin Biochem 2009;42:232-5.
2. Miller WG, Erek A, Cunningham TD, Oladipo O, Scott MG, Johnson RE. Commutability limitations influence quality control results with different reagent lots. Clin Chem 2011;57:76 – 83.
3. Sciacovelli L, Secchiero S, Zardo L, Zaninotto M, Plebani M. External Quality Assessment: an effective tool for Clinical Governance in laboratory medicine. Clin Chem Lab Med 2006;44:740-9.
4. Stepman HC, Stöckl D, Acheme R, Sesini S, Mazziotta D, Thienpont LM. Status of serum-calcium and -albumin measurement in Argentina assessed in 300 representative laboratories with 20 fresh frozen single donation sera. Clin Chem Lab Med 2011;49:1829-36.
5. Van Houcke SK, Rustad P, Stepman HC, Kristensen GB, Stöckl D, Røraas TH et al. Calcium, magnesium, albumin, and total protein measurement in serum as assessed with 20 fresh-frozen single-donation sera. Clin Chem 2012;58:1597-9.
6. Stepman HC, Tiikkainen U, Stöckl D, Vesper HW, Edwards SH, Laitinen H et al. Measurements for 8 common analytes in native sera identifies inadequate standardization among 6 routine laboratory assays. Clin Chem 2014;60:855-63.
7. CLSI. Preparation and validation of commutable frozen human serum pools as secondary reference materials for cholesterol measurement procedures; approved guideline. CLSI document C37-A. Wayne (PA): Clinical and Laboratory Standards Institute; 1999.
8. Westgard QC. Biological variation database, and quality specifications for imprecision, bias and total error (desirable and minimum). The 2014 update. http://www.westgard.com/biodatabase-2014-update.htm (accessed September 2014).
9. Nordic Reference Interval Project (NORIP). http://pweb.furst.no/norip/ (accessed September 2014).
10. STT-Consulting - Empower – Master Comparison 2014. http://www.stt-consulting.com/news.php?rubriek=8 / (accessed September 2014).
11. Schumann G, Bonora R, Ceriotti F, Clerc-Renaud P, Ferrero CA, Férard G 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;40:643-48.

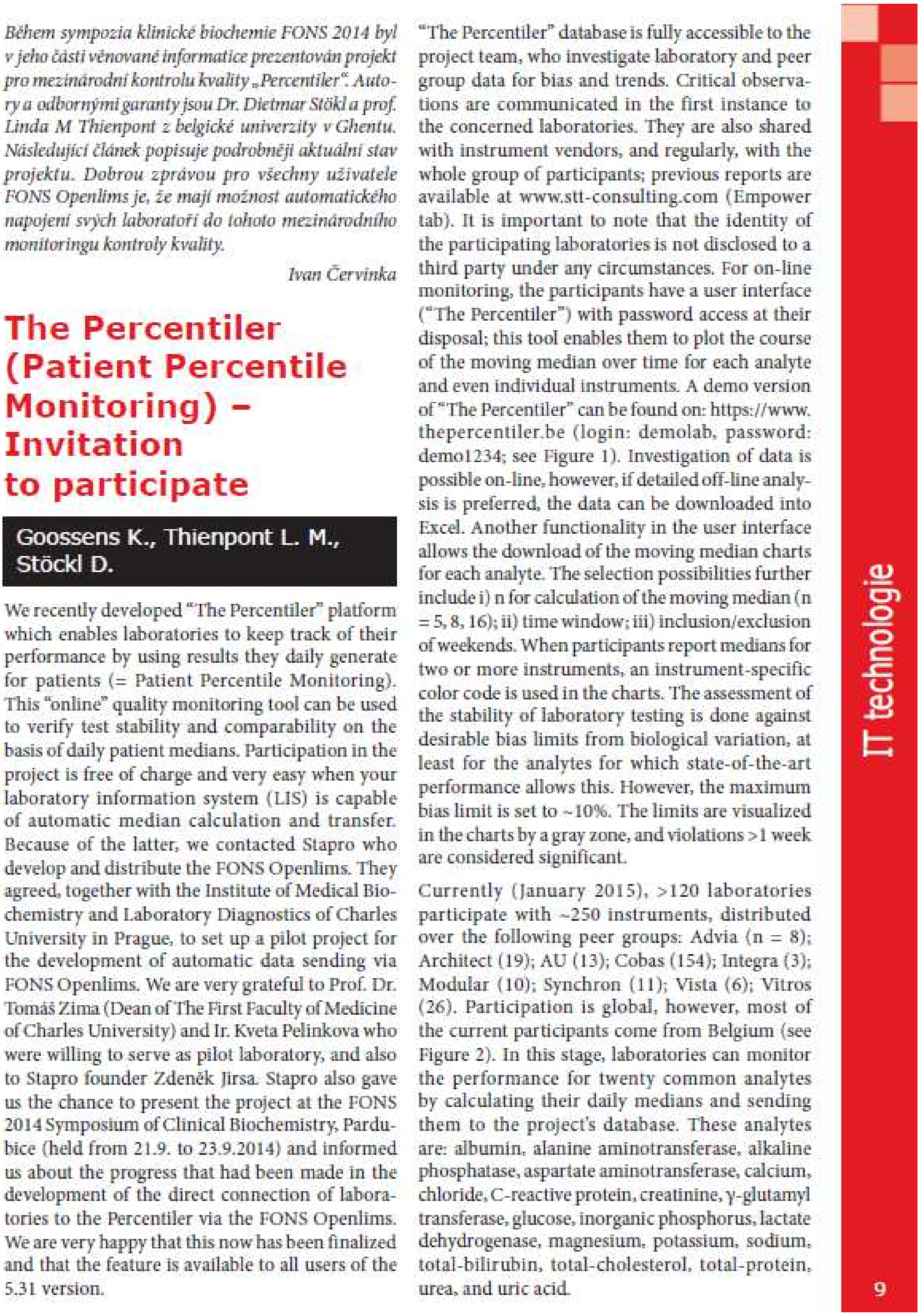
12. Schumann G, Bonora R, Ceriotti F, Férard G, Franck PFH, Gella F-J et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. Part 4. Reference procedure for the measurement of catalytic concentration of alanine aminotransferase. Clin Chem Lab Med 2002;40:718-24.
13. Schumann G, Bonora R, Ceriotti F, Férard G, Ferrero CA, Franck PFH et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. Part 5. Reference procedure for the measurement of catalytic concentration of aspartate aminotransferase. Clin Chem Lab Med 2002;40:725-33.
14. Schumann G, Bonora R, Ceriotti F, Férard G, Ferrero CA, Franck PFH et al. IFCC Primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. Part 6. Reference procedure for the measurement of catalytic concentration of γ-glutamyltransferase. Clin Chem Lab Med 2002;40: 734-38.
15. Schumann G, Klauke R, Canalias F, Bossert-Reuther S, Franck PFH, Gella F-J et al. IFCC Primary Reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C Part 9. Reference procedure for the measurement of catalytic concentration of alkaline phosphatase. Clin Chem Lab Med 2011;49:1439-46.
16. Stöckl D, Thienpont LM. The combined-target approach: a way out of the proficiency testing dilemma. Arch Pathol Lab Med. 1994;118:775-6.
17. Horowitz GL. Assessing accuracy on the front lines: a pragmatic approach for single-donor proficiency testing. Clin Chem 2014;60,806-8.
18. Van Houcke SK, Stepman HC, Thienpont LM, Fiers T, Stove V, Couck P et al. Long-term stability of laboratory tests and practical implications for quality management. Clin Chem Lab Med 2013;51:1227-31.
19. Tonks DB. A study of the accuracy and precision of clinical chemistry determinations in 170 Canadian laboratories. Clin Chem. 1963;9:217-33.
20. Kallner A, McQueen M, Heuck C. The Stockholm Consensus Conference on quality specifications in laboratory medicine, 25-26 April 1999. Scand J Clin Lab Invest 1999;59:475-6.
21. IFCC Working group on allowable errors for traceable results (WG-AETR). http://www.ifcc.org/ifcc-scientific-division/sd-working-groups/allowable-errors-for-traceable-results-wg-aetr/ (accessed September 2014).
22. 1st EFLM Strategic Conference - Defining analytical performance goals - 15 years after the Stockholm Conference. http://www.efcclm.eu/files/efcc/Leaflet

%20EFLM%20strategic%20conference.pdf (accessed September 2014).

23. Stepman HCM, Stöckl D, Stove V, Fiers T, Couck P, Gorus F, Thienpont LM. Long-term stability of clinical laboratory data - Sodium as benchmark. Clin Chem 2011;57:1616-7.
24. Arbor Research Collaborative for Health. Dialysis Outcomes and Practice Patterns Study. http://www.dopps.org (accessed September 2014).

Annex to chapter 5

Extract from “The Percentiler (Patient Percentile Monitoring) – Invitation to participate” (FONS bulletin):



Během sympozia klinické biochemie FONS 2014 byl v jeho části věnované informatice prezentován projekt pro mezinárodní kontrolu kvality „Percentiler“. Autory a odbornými garanty jsou Dr. Dietmar Stökl a prof. Linda M Thienpont z belgické univerzity v Ghentu. Následující článek popisuje podrobněji aktuální stav projektu. Dobrou zprávou pro všechny uživatele FONS Openlims je, že mají možnost automatického napojení svých laboratoří do tohoto mezinárodního monitoringu kontroly kvality.

Ivan Červinka

The Percentiler (Patient Percentile Monitoring) – Invitation to participate

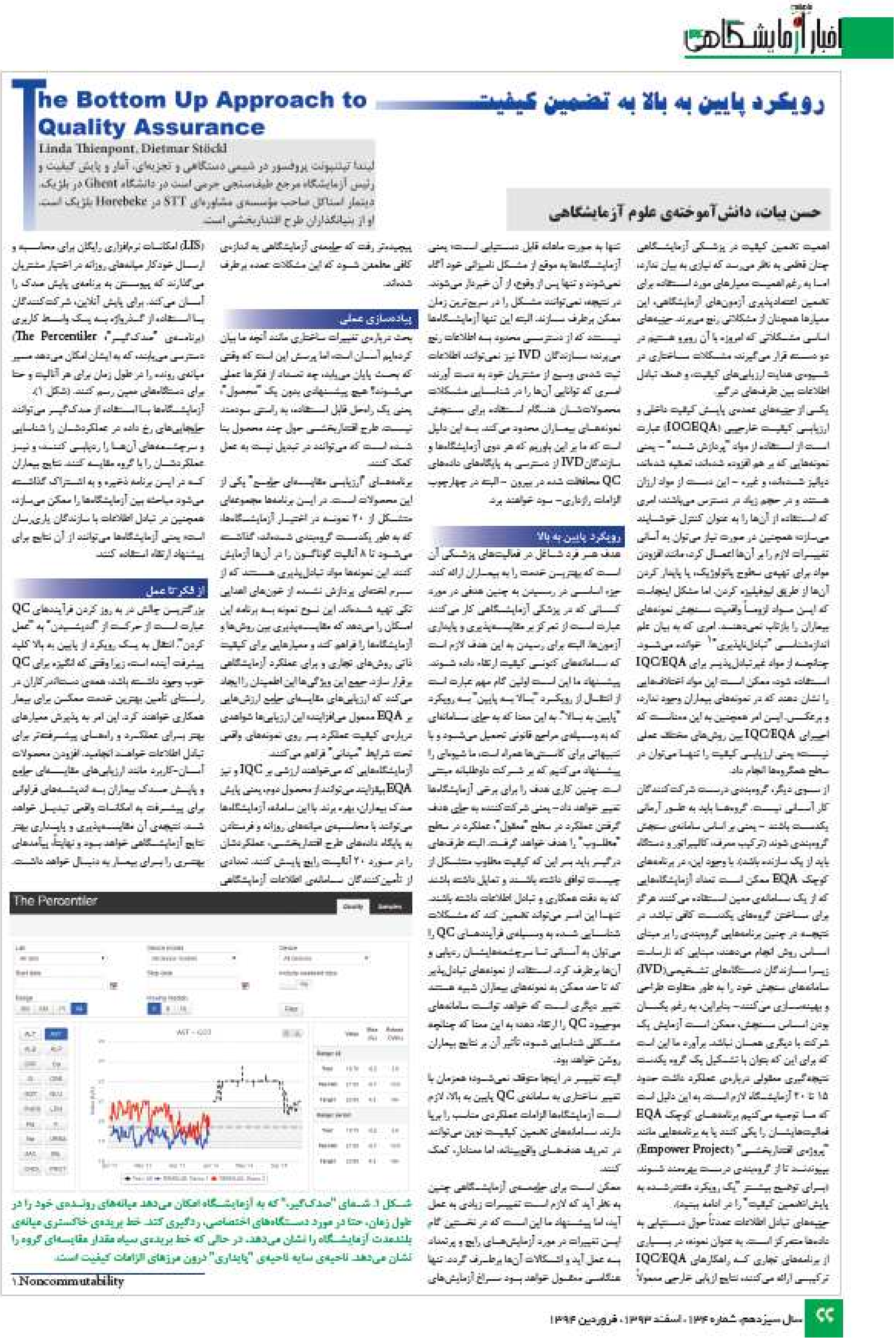
Goossens K., Thienpont L. M., Stöckl D.

We recently developed “The Percentiler” platform which enables laboratories to keep track of their performance by using results they daily generate for patients (= Patient Percentile Monitoring). This “online” quality monitoring tool can be used to verify test stability and comparability on the basis of daily patient medians. Participation in the project is free of charge and very easy when your laboratory information system (LIS) is capable of automatic median calculation and transfer. Because of the latter, we contacted Stapro who develop and distribute the FONS Openlims. They agreed, together with the Institute of Medical Biochemistry and Laboratory Diagnostics of Charles University in Prague, to set up a pilot project for the development of automatic data sending via FONS Openlims. We are very grateful to Prof. Dr. Tomáš Zima (Dean of The First Faculty of Medicine of Charles University) and Ir. Kveta Pelinkova who were willing to serve as pilot laboratory, and also to Stapro founder Zdeněk Jirsa. Stapro also gave us the chance to present the project at the FONS 2014 Symposium of Clinical Biochemistry, Pardubice (held from 21.9. to 23.9.2014) and informed us about the progress that had been made in the development of the direct connection of laboratories to the Percentiler via the FONS Openlims. We are very happy that this now has been finalized and that the feature is available to all users of the 5.31 version.

“The Percentiler” database is fully accessible to the project team, who investigate laboratory and peer group data for bias and trends. Critical observations are communicated in the first instance to the concerned laboratories. They are also shared with instrument vendors, and regularly, with the whole group of participants; previous reports are available at www.stt-consulting.com (Empower tab). It is important to note that the identity of the participating laboratories is not disclosed to a third party under any circumstances. For on-line monitoring, the participants have a user interface (“The Percentiler”) with password access at their disposal; this tool enables them to plot the course of the moving median over time for each analyte and even individual instruments. A demo version of “The Percentiler” can be found on: https://www.thepercentiler.be (login: demolab, password: demo1234; see Figure 1). Investigation of data is possible on-line, however, if detailed off-line analysis is preferred, the data can be downloaded into Excel. Another functionality in the user interface allows the download of the moving median charts for each analyte. The selection possibilities further include i) n for calculation of the moving median (n = 5, 8, 16); ii) time window; iii) inclusion/exclusion of weekends. When participants report medians for two or more instruments, an instrument-specific color code is used in the charts. The assessment of the stability of laboratory testing is done against desirable bias limits from biological variation, at least for the analytes for which state-of-the-art performance allows this. However, the maximum bias limit is set to ~10%. The limits are visualized in the charts by a gray zone, and violations >1 week are considered significant.

Currently (January 2015), >120 laboratories participate with ~250 instruments, distributed over the following peer groups: Advia (n = 8); Architect (19); AU (13); Cobas (154); Integra (3); Modular (10); Synchron (11); Vista (6); Vitros (26). Participation is global, however, most of the current participants come from Belgium (see Figure 2). In this stage, laboratories can monitor the performance for twenty common analytes by calculating their daily medians and sending them to the project's database. These analytes are: albumin, alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase, calcium, chloride, C-reactive protein, creatinine, γ-glutamyl transferase, glucose, inorganic phosphorus, lactate dehydrogenase, magnesium, potassium, sodium, total-bilirubin, total-cholesterol, total-protein, urea, and uric acid.

Extract from the Farsi translation of “The Bottom Up Approach to Quality Assurance”



اخبار آزمایشگاهی

رویکرد پایین به بالا به تضمین کیفیت

The Bottom Up Approach to Quality Assurance

Linda Thienpont, Dietmar Stöckl

لیندا تینپونت پروفسور در شیمی دستگاهی و تجزیه‌ای، آمار و پایش کیفیت و رئیس آزمایشگاه مرجع طیف‌سنجی جرمی است در دانشگاه Ghent در بلژیک.
دیتمار استاکل صاحب مؤسسه‌ی مشاوره‌ای STT در Horebeke بلژیک است.
او از بنیانگذاران طرح اقتداربخشی است.

حسن بیات، دانش‌آموخته‌ی علوم آزمایشگاهی

اهمیت تضمین کیفیت در پزشکی آزمایشگاهی چنان قطعی به نظر می‌رسد که نیازی به بیان ندارد، اما به رغم اهمیت معیارهای مورد استفاده برای تضمین اعتمادپذیری آزمون‌های آزمایشگاهی، این معیارها همچنان از مشکلاتی رنج می‌برند. جنبه‌های اساسی مشکلاتی که امروزه با آن روبرو هستیم در دو دسته قرار می‌گیرند: مشکلات ساختاری در شیوه‌ی هدایت ارزیابی‌های کیفیت، و ضعف تبادل اطلاعات بین طرف‌های درگیر.

یکی از جنبه‌های عمده‌ی پایش کیفیت داخلی و ارزیابی کیفیت خارجی (IQC/EQA) عبارت است از استفاده از مواد "پردازش شده" – یعنی نمونه‌هایی که بر هم افزوده شده‌اند، تصفیه شده‌اند، دیالیز شده‌اند، و غیره – این دسته از مواد ارزان هستند و در حجم زیاد در دسترس می‌باشند. امری که استفاده از آن‌ها را به عنوان کنترل خوشایند می‌سازد. همچنین در صورت نیاز می‌توان به آسانی تغییرات لازم را بر آن‌ها اعمال کرد، مانند افزودن مواد برای تهیه‌ی سطوح پاتولوژیک، یا پایدار کردن آن‌ها از طریق لیوفلیزه کردن. اما مشکل اینجاست که این مواد لزوماً واقعیت سنجش نمونه‌های بیماران را بازتاب نمی‌دهند. امری که به بیان علم اندازه‌شناسی "تبادل‌ناپذیری"[1] خوانده می‌شود. چنانچه از مواد غیرتبادل‌پذیر برای IQC/EQA استفاده شود، ممکن است این مواد اختلاف‌هایی را نشان دهند که در نمونه‌های بیماران وجود ندارد، و برعکس. این امر همچنین به این معناست که اجرای IQC/EQA بین روش‌های مختلف عملی نیست، یعنی ارزیابی کیفیت را تنها می‌توان در سطح همگروه‌ها انجام داد.

از سوی دیگر، گروه‌بندی درست شرکت‌کنندگان کار آسانی نیست. گروه‌ها باید به طور آرمانی یکدست باشند – یعنی بر اساس سامانه‌ی سنجش گروه‌بندی شوند (ترکیب معرف، کالیبراتور و دستگاه باید از یک سازنده باشند). با وجود این، در برنامه‌های کوچک EQA ممکن است تعداد آزمایشگاه‌هایی که از یک سامانه‌ی معین استفاده می‌کنند هرگز برای ساختن گروه‌های یکدست کافی نباشد. در نتیجه در چنین برنامه‌هایی گروه‌بندی را بر مبنای اساس روش انجام می‌دهند، مبنایی که نارساست زیرا سازندگان دستگاه‌های تشخیصی (IVD) سامانه‌های سنجش خود را به طور متفاوت طراحی و بهینه‌سازی می‌کنند – بنابراین، به رغم یکسان بودن اساس سنجش، ممکن است آزمایش یک شرکت با دیگری همسان نباشد. برآورد ما این است که برای این که بتوان با تشکیل یک گروه یکدست نتیجه‌گیری معقولی درباره‌ی عملکرد داشت حدود ۱۵ تا ۲۰ آزمایشگاه لازم است. به این دلیل است که ما توصیه می‌کنیم برنامه‌های کوچک EQA فعالیت‌هایشان را یکی کنند یا به برنامه‌هایی مانند "پروژه‌ی اقتداربخشی" (Empower Project) بپیوندند تا از گروه‌بندی درست بهره‌مند شوند (برای توضیح بیشتر "یک رویکرد مقتدرشده به پایش/تضمین کیفیت" را در ادامه ببینید).

جنبه‌های تبادل اطلاعات عمدتاً حول دستیابی به داده‌ها متمرکز است. به عنوان نمونه، در بسیاری از برنامه‌های تجاری که راهکارهای IQC/EQA ترکیبی ارائه می‌کنند، نتایج ارزیابی خارجی معمولاً تنها به صورت ماهانه قابل دستیابی است؛ یعنی آزمایشگاه‌ها به موقع از مشکل نامیزانی خود آگاه نمی‌شوند و تنها پس از وقوع، از آن خبردار می‌شوند. در نتیجه، نمی‌توانند مشکل را در سریع‌ترین زمان ممکن برطرف سازند. البته این تنها آزمایشگاه‌ها نیستند که از دسترسی محدود به اطلاعات رنج می‌برند. سازندگان IVD نیز نمی‌توانند اطلاعات ثبت شده‌ی وسیع از مشتریان خود به دست آورند. امری که توانایی آن‌ها را در شناسایی مشکلات محصولات‌شان هنگام استفاده برای سنجش نمونه‌های بیماران محدود می‌کند. به این دلیل است که ما بر این باوریم که هر دوی آزمایشگاه‌ها و سازندگان IVD از دسترسی به پایگاه‌های داده‌های QC محافظت شده در بیرون – البته در چهارچوب الزامات رازداری – سود خواهند برد.

رویکرد پایین به بالا

هدف هر فرد شاغل در فعالیت‌های پزشکی آن است که بهترین خدمت را به بیماران ارائه کند. جزء اساسی در رسیدن به چنین هدفی در مورد کسانی که در پزشکی آزمایشگاهی کار می‌کنند عبارت است از تمرکز بر مقایسه‌پذیری و پایداری آزمون‌ها. البته برای رسیدن به این هدف لازم است که سامانه‌های کنونی کیفیت ارتقاء داده شوند. پیشنهاد ما این است اولین گام مهم عبارت است از انتقال از رویکرد "بالا به پایین" به رویکرد "پایین به بالا". به این معنا که به جای سامانه‌هایی که به وسیله‌ی مراجع قانونی تحمیل می‌شود و با تنبیهاتی برای کاستی‌ها همراه است، ما شیوه‌ای را پیشنهاد می‌کنیم که بر شرکت داوطلبانه مبتنی است. چنین کاری هدف را برای برخی آزمایشگاه‌ها تغییر خواهد داد – یعنی شرکت‌کننده به جای هدف گرفتن عملکرد در سطح "معقول"، عملکرد در سطح "مطلوب" را هدف خواهد گرفت. البته طرف‌های درگیر باید بر این که کیفیت مطلوب متشکل از چیست توافق داشته باشند و تمایل داشته باشند که به دقت همکاری و تبادل اطلاعات داشته باشند. تنها این امر می‌تواند تضمین کند که مشکلات شناسایی شده به وسیله‌ی فرآیندهای QC را می‌توان به آسانی تا سرچشمه‌هایشان ردیابی و آن‌ها برطرف کرد. استفاده از نمونه‌های تبادل‌پذیر که تا حد ممکن به نمونه‌های بیماران شبیه هستند تغییر دیگری است که خواهد توانست سامانه‌های موجود QC را ارتقاء دهد؛ به این معنا که چنانچه مشکلی شناسایی شود، تاثیر آن بر نتایج بیماران روشن خواهد بود.

البته تغییر در اینجا متوقف نمی‌شود؛ همزمان با تغییر ساختاری به سامانه‌ی QC پایین به بالا، لازم است آزمایشگاه‌ها الزامات عملکردی مناسب را برپا دارند. سامانه‌های تضمین کیفیت نوین می‌توانند در تعریف هدف‌های واقع‌بینانه، اما معنادار، کمک کنند.

ممکن است برای جامعه‌ی آزمایشگاهی چنین به نظر آید که لازم است تغییرات زیادی به عمل آید، اما پیشنهاد ما این است که در نخستین گام این تغییرات در مورد آزمایش‌های رایج و پرتعداد به عمل آید و اشکالات آن‌ها برطرف گردد. تنها هنگامی معقول خواهد بود سراغ آزمایش‌های پیچیده‌تر رفت که جامعه‌ی آزمایشگاهی به اندازه‌ی کافی مطمئن شود که این مشکلات عمده برطرف شده‌اند.

پیاده‌سازی عملی

بحث درباره‌ی تغییرات ساختاری مانند آنچه ما بیان کردیم آسان است، اما پرسش این است که وقتی که بحث پایان می‌یابد، چه تعداد از فکرها عملی می‌شوند؟ هیچ پیشنهادی بدون یک "محصول"، یعنی یک راه‌حل قابل استفاده، به راستی سودمند نیست. طرح اقتداربخشی حول چند محصول بنا شده است که می‌توانند در تبدیل نیت به عمل کمک کنند.

برنامه‌های "ارزیابی مقایسه‌ای جامع" یکی از این محصولات است. در این برنامه‌ها مجموعه‌های متشکل از ۲۰ نمونه در اختیار آزمایشگاه‌ها، که به طور یکدست گروه‌بندی شده‌اند، گذاشته می‌شود تا ۸ آنالیت گوناگون را در آن‌ها آزمایش کنند. این نمونه‌ها مواد تبادل‌پذیری هستند که از سرم لخته‌ی پردازش نشده از خون‌های اهدایی تکی تهیه شده‌اند. این نوع نمونه به برنامه این امکان را می‌دهد که مقایسه‌پذیری بین روش‌ها و آزمایشگاه‌ها را فراهم کند و معیارهایی برای کیفیت ذاتی روش‌های تجاری و برای عملکرد آزمایشگاهی برقرار سازد. جمع این ویژگی‌ها این اطمینان را ایجاد می‌کند که ارزیابی‌های مقایسه‌ای جامع ارزش‌هایی بر EQA معمول می‌افزایند. این ارزیابی‌ها شواهدی درباره‌ی کیفیت عملکرد بر روی نمونه‌های واقعی تحت شرایط "میدانی" فراهم می‌کنند.

آزمایشگاه‌هایی که می‌خواهند ارزشی بر IQC و نیز EQA بیفزایند می‌توانند از محصول دوم، یعنی پایش صدک بیماران، بهره برند. با این سامانه، آزمایشگاه‌ها می‌توانند با محاسبه‌ی میانه‌های روزانه و فرستادن به پایگاه داده‌های طرح اقتداربخشی، عملکردشان را در مورد ۲۰ آنالیت رایج پایش کنند. تعدادی از تأمین‌کنندگان سامانه‌ی اطلاعات آزمایشگاهی (LIS) امکانات نرم‌افزاری رایگان برای محاسبه و ارسال خودکار میانه‌های روزانه در اختیار مشتریان می‌گذارند که پیوستن به برنامه‌ی پایش صدک را آسان می‌کند. برای پایش آنلاین، شرکت‌کنندگان با استفاده از گذرواژه، به یک واسط کاربری (برنامه‌ی "صدک‌گیر"؛ The Percentiler) دسترسی می‌یابند، که به ایشان امکان می‌دهد مسیر میانه‌ی رونده را در طول زمان برای هر آنالیت و حتا برای دستگاه‌های معین رسم کنند. (شکل ۱).

آزمایشگاه‌ها با استفاده از صدک‌گیر می‌توانند جابجایی‌های رخ داده در عملکردشان را شناسایی و سرچشمه‌های آن‌ها را ردیابی کنند و نیز عملکردشان را با گروه مقایسه کنند. نتایج بیماران که در این برنامه ذخیره و به اشتراک گذاشته می‌شود مباحثه بین آزمایشگاه‌ها را ممکن می‌سازد، همچنین در تبادل اطلاعات با سازندگان یاری‌رسان است؛ یعنی آزمایشگاه‌ها می‌توانند از آن نتایج برای پیشنهاد ارتقاء استفاده کنند.

از فکر تا عمل

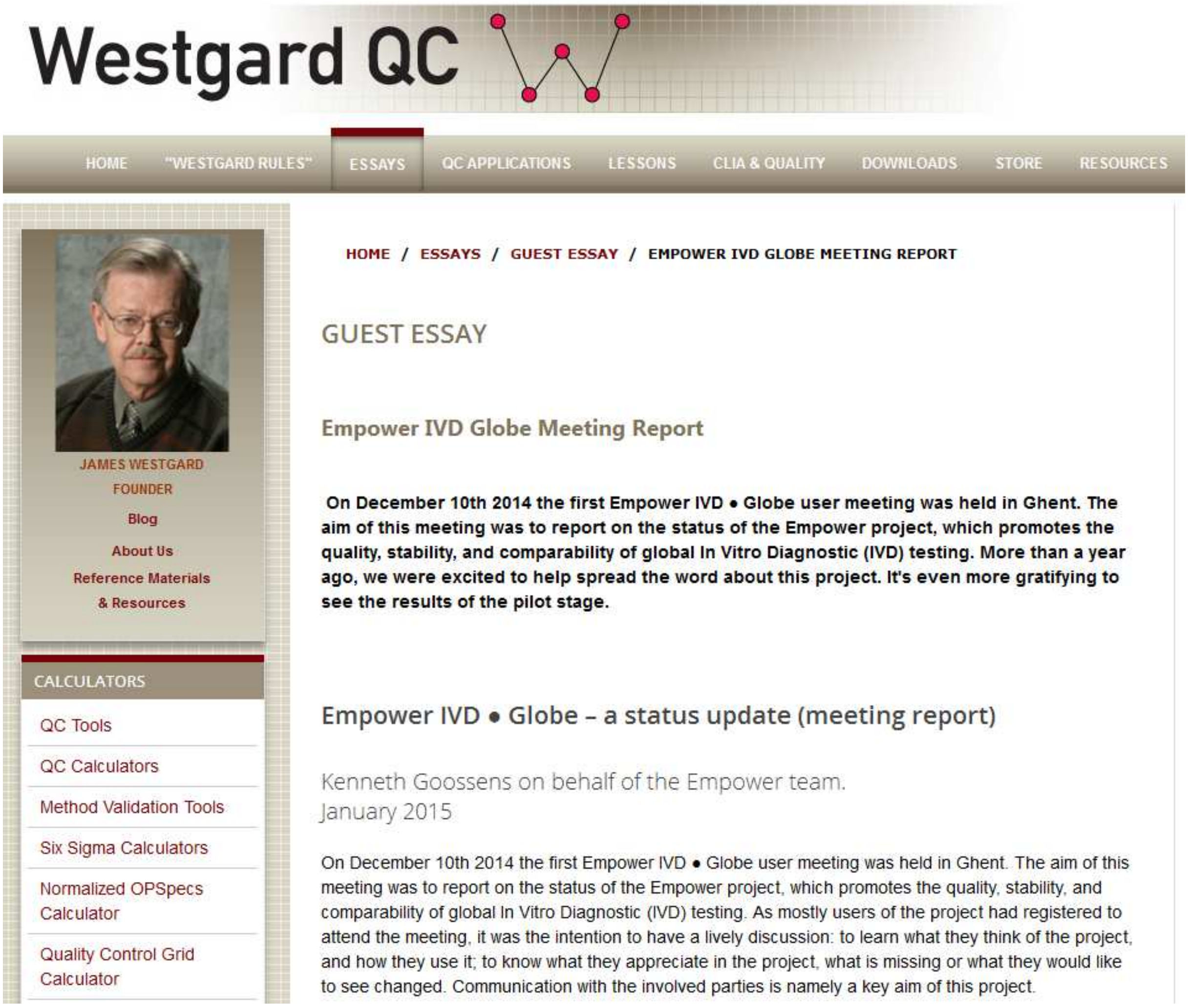
بزرگترین چالش در به روز کردن فرآیندهای QC عبارت است از حرکت از "اندیشیدن" به "عمل کردن". انتقال به یک رویکرد از پایین به بالا کلید پیشرفت آینده است، زیرا وقتی که انگیزه برای QC خوب وجود داشته باشد، همه‌ی دست‌اندرکاران در راستای تأمین بهترین خدمت ممکن برای بیمار همکاری خواهند کرد. این امر به پذیرش معیارهای بهتر برای عملکرد و راه‌های پیشرفته‌تر برای تبادل اطلاعات خواهد انجامید. افزودن محصولات آسان-کاربرد مانند ارزیابی‌های مقایسه‌ای جامع و پایش صدک بیماران به اندیشه‌های فراوانی برای پیشرفت به امکانات واقعی تبدیل خواهد شد. نتیجه‌ی آن مقایسه‌پذیری و پایداری بهتر نتایج آزمایشگاهی خواهد بود و نهایتاً پیامدهای بهتری را برای بیمار به دنبال خواهد داشت.

شکل ۱. شمای "صدک‌گیر" که به آزمایشگاه امکان می‌دهد میانه‌های رونده‌ی خود را در طول زمان، حتا در مورد دستگاه‌های اختصاصی، ردگیری کند. خط بریده‌ی خاکستری میانه‌ی بلندمدت آزمایشگاه را نشان می‌دهد، در حالی که خط بریده‌ی سیاه مقدار مقایسه‌ای گروه را نشان می‌دهد. ناحیه‌ی سایه ناحیه‌ی "پایداری" درون مرزهای الزامات کیفیت است.

[1] Noncommutability

Extract from the report of the Empower IVD●Globe meeting – Guest assay on Westgard QC

https://www.westgard.com/empower-ivd-globe-2014.htm



GUEST ESSAY

Empower IVD Globe Meeting Report

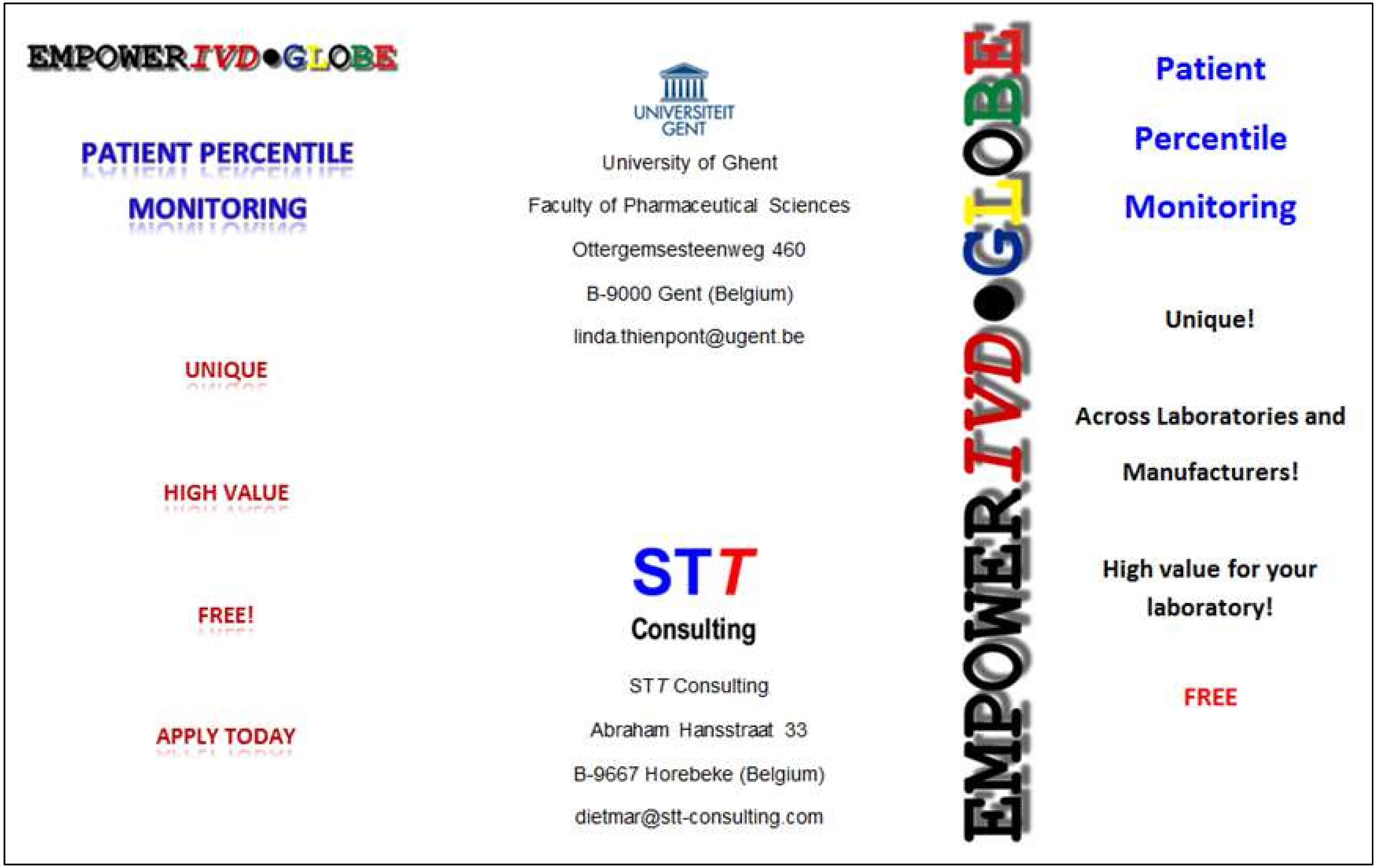
On December 10th 2014 the first Empower IVD • Globe user meeting was held in Ghent. The aim of this meeting was to report on the status of the Empower project, which promotes the quality, stability, and comparability of global In Vitro Diagnostic (IVD) testing. More than a year ago, we were excited to help spread the word about this project. It's even more gratifying to see the results of the pilot stage.

Empower IVD • Globe – a status update (meeting report)

Kenneth Goossens on behalf of the Empower team.
January 2015

On December 10th 2014 the first Empower IVD • Globe user meeting was held in Ghent. The aim of this meeting was to report on the status of the Empower project, which promotes the quality, stability, and comparability of global In Vitro Diagnostic (IVD) testing. As mostly users of the project had registered to attend the meeting, it was the intention to have a lively discussion: to learn what they think of the project, and how they use it; to know what they appreciate in the project, what is missing or what they would like to see changed. Communication with the involved parties is namely a key aim of this project.

Extract from the commercial flyer distributed for participation to the Percentiler and Flagger



EMPOWER*IVD*●GLOBE

PATIENT PERCENTILE

MONITORING

UNIQUE

HIGH VALUE

FREE!

APPLY TODAY

UNIVERSITEIT
GENT

University of Ghent
Faculty of Pharmaceutical Sciences
Ottergemsesteenweg 460
B-9000 Gent (Belgium)
linda.thienpont@ugent.be

ST*T*
Consulting

ST*T* Consulting
Abraham Hansstraat 33
B-9667 Horebeke (Belgium)
dietmar@stt-consulting.com

EMPOWER*IVD*●GLOBE

Patient

Percentile

Monitoring

Unique!

Across Laboratories and

Manufacturers!

High value for your
laboratory!

FREE

6 Chapter VI

Statistical analysis from a different viewpoint

This chapter is based upon:

- Goossens K, De Grande L, Keller T, Weber S, Thienpont L. Verification of reference intervals by the C28 protocol – The alpha error/power trade-off. Clin Chim Acta 2014;436:18–19.
- Goossens K, Van Uytfanghe K, Thienpont L. Calculation of non-commutability budgets by t-testing. Clin Chim Acta 2015;438:212–213.

Highlights

The appropriate use of statistics is essential for correct data interpretation in the clinical laboratory. Several guidelines are available which aid clinical chemists for their local data processing and interpretation, e.g. guidelines for reference interval validation and guidelines for commutability assessment. These guidelines, however, have some limitations. That's why we decided to discuss the issues for two of these guidelines and propose a valid alternative.

Highlights of the research:

- The current C28 reference interval validation protocol should be handled with caution. Optimized sampling protocols with an increased number of samples might be recommendable to achieve more statistical power.
- Current commutability assessment protocols fail to recognize non-commutability when the analytical variation of a measurement method is quite high.
- Our t-testing approach can be used as an alternative for commutability assessment with consideration of a desirable non-commutability budget in relation to the sample size.

1. Introduction

As data handling in the clinical environment heavily depends on the use of statistics, clinical chemists face the need to analyze data using the most appropriate statistical tools. On the one hand the clinical laboratory relies on descriptive statistics, where it needs to quantitatively or graphically describe the data. On the other hand, it relies on inferential statistics where it makes statements about the population based on a selection of samples. Data interpretation also depends on the nature of the data (e.g. qualitative or quantitative data, discrete or continuous data) and each data category requires a different approach for analysis. To aid the clinical chemist with selecting the correct statistical approach, the CLSI provides several guidelines on a range of different topics (1). These guidelines, however, are not without limitations. In this chapter, we will briefly discuss two of those guidelines, point to the limitations which can occur, and provide an alternative to avoid them. First, we will provide an alternative for the double sampling protocol from the C28 guidelines and secondly for the current commutability assessment protocols from the EP14 and EP30 guidelines.

2. Verifying Reference intervals in the clinical laboratory

The C28 guideline from CLSI provides laboratories and manufacturers with specific recommendations for procedures which can be applied to establish and verify reference intervals for quantitative laboratory tests (2). By using this document laboratories are, amongst others, able to verify the appropriateness of reference intervals for their own population. The laboratories can assess the validity of existing reference limits by examining a small number of reference individuals (n = 20) with a binomial test. This assessment is performed with a double sampling protocol. When no more than 2 of the 20 tested individuals fall outside the original reference interval, the original limits are considered valid. When 3 or 4 results fall outside the original reference interval, a second sampling is performed with, again, 20 reference individuals. Similar to the first sampling, the limits are considered valid when no more than 2 test results fall outside the original reference interval. When 3 or more results fall outside the limits in the second sampling set, or when 5 or more results fall outside the limits in the first set, the laboratory is advised to consider developing its own reference interval.

This double sampling protocol is applied because it significantly lowers the chance of falsely rejecting the existing reference interval. The alpha error equals 7.5%, when the reference limits are rejected if more than two test results fall outside the reference interval, and it drops from 7.5% to less than 1% when the double sampling protocol is used. However, at the same time, using this approach will lower the power of the test, thereby increasing the chance of falsely accepting the existing reference interval. This trade-off between alpha error and power is discussed in the letter to editor on pages 231-233, where we also describe an alternative in the form of an increased number of samples, i.e. 100 samples instead of 20 (3).

3. Commutability assessment of reference materials

As mentioned in the introduction, traceability of laboratory measurements is essential to create comparable results irrespective of time, location, and the measurement system which was used (4). Reference materials play an important role in the traceability chain because they are used to link patient results to the accepted reference. However, certified reference materials used for method calibration can only provide metrological traceability of a measurement procedure when they are commutable with it (5,6). This means that they should mimic patient samples as closely as possible to minimize matrix effects.

According to the EP30-A CLSI document, commutability is defined as “equivalence of mathematical relationship among the results of different measurement procedures targeting the same measurement for a reference material and for representative samples of the type intended to be measured” (7). In practice this means, when both reference materials and clinical samples are measured with two or more analytical methods, the reference materials can only be considered as commutable when they have interassay properties similar to those of the clinical samples. If a reference material is not commutable, e.g. due to matrix effects, it can’t be used as a calibrator, or better said it will not result in traceability of that result. Absence of commutability would introduce an artificial bias, thereby making it impossible to determine whether an analytical bias exists in calibration. Several statistical analyses have been described to assess commutability of a candidate reference material. Current approaches investigate whether the reference material is part of the population of clinical samples at a certain probability. For example, whether the reference material is within the 95% confidence ellipse of multivariate statistics of the clinical samples (8), within the ± $3S_{y/x}$ (or another multiple of the standard error of regression) range around the mean (9), or within the 95% prediction interval around the regression line (10). An extensive overview of these methods can be found in the review by Vesper *et al* (11). The aforementioned approach, based on prediction intervals, is adopted by the CLSI EP14 guideline (12). This protocol states that commutability is considered likely when the mean value for the reference material falls inside the prediction interval around the regression line constructed with the clinical samples (under the condition that linear regression analysis is suitable). Usually, 20 clinical samples are measured in trifold in order to construct the regression line. The width of the prediction interval around that regression line is inversely proportional to the accuracy of the applied procedure. When the accuracy is low, the prediction interval will have a high width, when the accuracy is high a low width. This means that methods with lower accuracy are less likely to detect a non-commutable reference material.

In our Clinica Chimica Acta manuscript (13), we describe a new approach that investigates whether the reference material is equivalent to the average clinical sample. It is based on the two-one-sided t-test approach (TOST, 14). Whereas in literature clinical samples are chosen within a wide concentration range to cover the entire clinically relevant area, the experiment described in our TOST-approach uses a small range: all clinical samples are from a concentration range which is very close to the reference material value. The data of the experiment are analyzed using the differences of clinical samples as well as reference materials. First, the absolute difference between the mean difference of clinical samples measured by two measurement methods and the mean difference of reference materials measured by the same two measurement methods is calculated (|∆CS - ∆RM| in Figure 1). This value will be equal to or greater than zero. Then a 95%-confidence interval is calculated around this absolute difference. The size of this interval will depend on the number of replicates measured for both the reference material and the clinical samples. The higher the number of replicates, the easier it becomes to prove the equivalence of the reference material to the clinical samples. The confidence interval around the absolute difference is then compared to a certain test limit. In the manuscript we set these limits equal to the bias limit based on the biological variation model extracted from the Westgard Biodatabase (15). Only if the confidence interval is entirely within the interval indicated by the test limit, commutability of reference materials compared to clinical samples is shown at the 95% certainty level. In case of the confidence interval enclosing the test limit, or when the confidence interval is entirely located outside of the interval, commutability has to be denied (see figure 1). A more extensive explanation on our t-test approach can be found in article 8 (pages 234-236).

Note, our manuscript is not the only literature which provides an alternative method for commutability assessment. Also other researchers in the commutability domain are moving away from the traditional methods using prediction intervals, and are handling commutability assessment with a t-test approach (16).

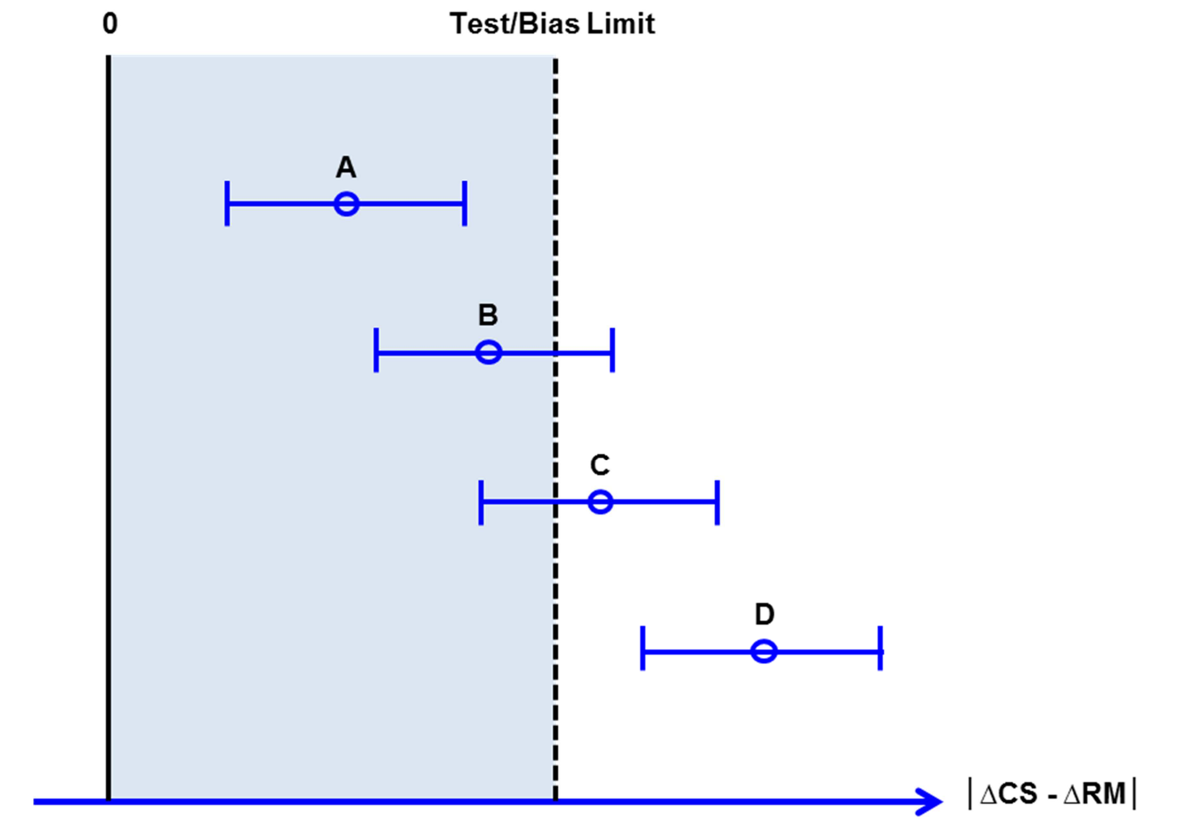


Figure 1. Interpretation of commutability of a reference material to the clinical samples.

A: The reference material is not significantly different from clinical sample because the two-sided 95% confidence interval doesn't cross the test limit.

B: The reference material is not significantly different from clinical sample but is not equivalent because part of the confidence interval lies outside the range indicated by the test limit. In this case a higher number of measurements could prove the commutability of the reference material.

C: The reference material is significantly different from clinical sample but not with 95% certainty.

D: The estimate is significantly different from control and is not equivalent to the control, because the 95% confidence interval is located completely outside the interval indicated with the test limit.

References

1. Clinical and Laboratory Standards Institute - Learn how you can improve laboratory testing quality. http://clsi.org/ (accessed July 2014).
2. Wayne, PA. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline-third edition. CLSI document C28-A3c. Clinical and Laboratory Standard Institute; 2008.
3. Goossens K, De Grande L, Keller T, Weber S, Thienpont L. Verification of reference intervals by the C28 protocol – The alpha error/power trade-off. Clin Chim Acta. 2014;436:18-9.
4. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55(6):1067-75.
5. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52(4):553-4.
6. Miller WG, Myers GL. Commutability still matters. Clin Chem 2013;59(9):1291-3.
7. Wayne, PA. CLSI. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline. CLSI document EP30-A. Clinical and Laboratory Standards Institute; 2010.
8. Rej R. Accurate enzyme activity measurements. Two decades of development in the commutability of enzyme quality control materials. Arch Pathol Lab Med. 1993;117:352–64.
9. Franzini C. Commutability of reference materials in clinical chemistry. J Int Fed Clin Chem. 1993;5:169–73.
10. Eckfeldt JH, Copeland KR. Accuracy verification and identification of matrix effects. Arch Pathol Lab Med. 1993;117:381–6.
11. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. Clin Biochem Rev 2007;28:139–47.
12. Wayne, PA. CLSI. Evaluation of commutability of processed samples; approved guideline-third edition. CLSI document EP14-A3. Clinical and Laboratory Standard Institute; 2014.
13. Goossens K, Van Uytfanghe K, Thienpont L. Calculation of non-commutability budgets by t-testing. Clin Chim Acta 2015;438:212–213.
14. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. J Pharmacokin Biopharm 1987;15:657–80.
15. Westgard QC – Biodatabase. http:// www.westgard.com/biodatabase1.htm (acessed July 2014).

16. Korzun WJ, Nilsson G, Bachmann LM, Myers G, Sakurabayashi I, Nakajima K, Nakamura M, Shamburek RD, Remaley AT, Miller WG. Difference in bias approach for commutability assessment: application to frozen pools of human serum measured by 8 direct methods for HDL and LDL cholesterol. Clin Chem 2015;61(8):1107-13.

Article 7. Verification of reference intervals by the C28 protocol (letter to editor)

Verification of reference intervals by the C28 protocol – The alpha error/power trade-off

Kenneth Goossens[1], Linde AC De Grande[1], Thomas Keller[2], Stefan Weber[2], Linda M Thienpont[1*]

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium; [2]ACOMED statistik, Leipzig, Germany

Medical laboratories are required to establish their own reference intervals or at least verify those published or given by the in-vitro diagnostic manufacturer. Typically, laboratories use the verification protocol recommended in the C28-guidelines of the Clinical and Laboratory Standard Institute [1]. It makes use of twenty reference individuals and applies a simple binomial test. A reference interval is considered valid when no more than two out of twenty samples fall outside the reference limits. If three or four samples are not included in the interval, a second sampling is recommended. Again, two samples are allowed to fall outside the reference limits. If more samples exceed the range, the laboratory should consider whether or not to develop its own reference interval.

According to the C28 guidelines the reason for performing the double sampling protocol is to lower the alpha error. Indeed, whilst single sampling results in an alpha error or false rejection of the reference interval under investigation in 7.5% of the cases, it drops to less than 1% with double sampling. However, the change in power (or beta error), which is defined as the probability to get a significant test, is not considered. While power curves are shown in the C28 guidelines for the single sampling, none are described for the protocol in its entirety. Also, the power curves shown in the C28 guidelines for the displacement of the mean are not using the SD units of the original distribution.

Here we present a more generalized form of these power curves and an alternative sampling protocol. We performed calculations with Microsoft Excel 2010 using the NORMSINV formula and calculated the power using the binomial formula. Figure 1 shows power curves for both the single (blue) and double (red) sampling protocol, limited to a Gaussian distribution and two-sided reference intervals. The x-axis shows the difference between the population distribution under investigation versus the reference population. It is expressed as a

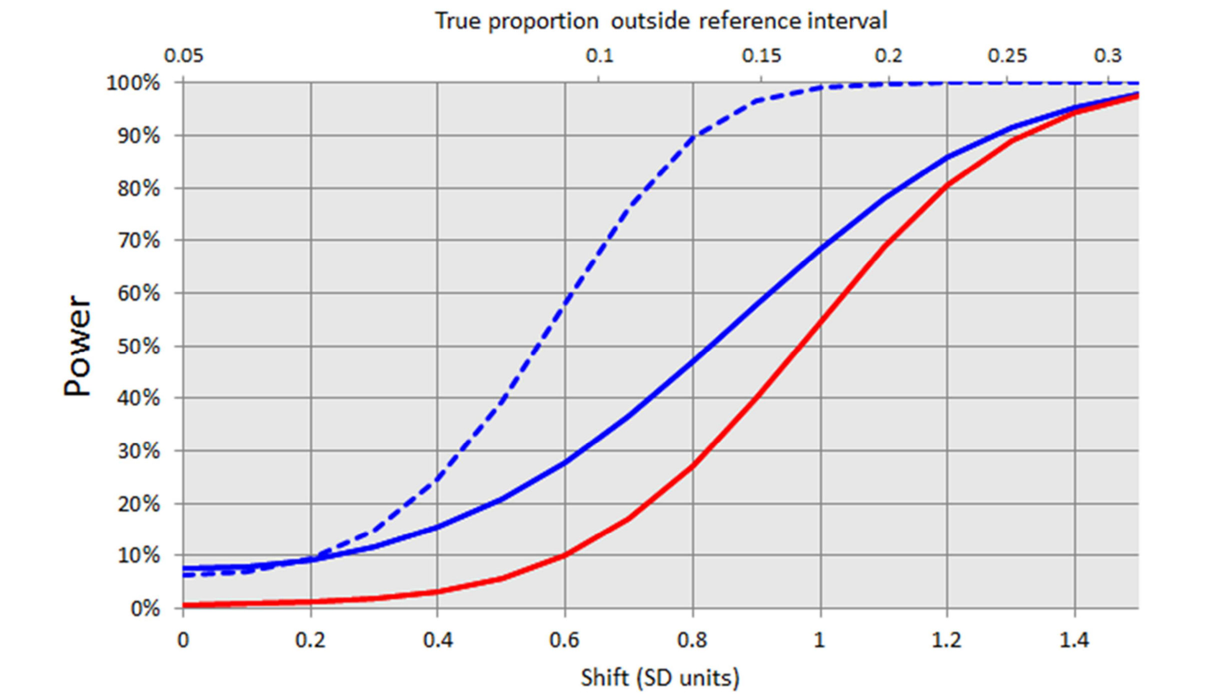


Figure 1: Power curves for the single and double sampling C28 protocol. Single sampling: two out of twenty samples allowed outside the reference interval (solid blue line), eight out of hundred samples (dotted blue line); double sampling (solid red line). The lower x-axis is expressed as a shift in SD units, the upper x-axis in true proportion outside the reference interval.

shift by 0 to 1.5 times the SD (relative to the reference population, assuming equal variances and two-tailed distribution). The third curve (blue – dotted line) shows a single sampling case with a higher sample number (allowing eight out of hundred samples to fall outside the reference interval). For completeness an alternative x-axis shows the true proportion outside the reference interval. In this case the assumption of normal distribution of data is not needed.

The power curves for the single and double sampling protocol show that, while the alpha error lowers, power is lost when using the double sampling protocol. Laboratories are more likely to assume the existing reference intervals are suited for use with their population, while they should have made the conclusion that a different reference interval is needed. This is a classic case of a trade-off between false positives and false negatives (or alpha error and power). It demonstrates that controlling for falsely positive assignments, without reporting beta errors, can be misleading.

Therefore, we propose to handle current validation protocols with caution. It might even be worthwhile to reconsider the C28 approach, and look for optimized sampling protocols. One possible approach could be to increase the sample size towards more statistical power (which was already suggested in the C28 protocol, e.g. using sixty or hundred samples instead of twenty). The third power curve in the graph illustrates that allowing eight samples out of hundred outside the reference interval results in a higher power, and still is capable of achieving a reasonable alpha error.

References

1. Wayne, PA. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline-third edition. CLSI document C28-A3c. Clinical and Laboratory Standard Institute; 2008.

Article 8. Non-commutability budgets (letter to editor).

Calculation of non-commutability budgets by t-testing.

Kenneth Goossens[1], Katleen Van Uytfanghe[1], Linda M Thienpont[1*]

[1]Laboratory for Analytical Chemistry, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium

Reference materials used in laboratory medicine either for method calibration or trueness control, need to be commutable. Commutability assessment is done by measurement of the reference material and clinical samples by at least two different analytical methods. Commutability is inferred from statistically testing the closeness of agreement between the mathematical relationship of the measurement results obtained by the methods for both groups of samples. A non-commutable reference material potentially causes artificial measurement biases, which can be attributed either to the presence of matrix effects or a non-native analyte. This leads necessarily to calibration errors or false conclusions about the (dis)agreement of results for clinical samples among methods [1]. Notwithstanding this, we think it is useful to define a reasonable budget (= non relevant difference between the mean concentrations for the reference material and clinical samples) by which a reference material can be tolerated to be commutable.

Current approaches for commutability testing investigate whether the reference material is part of the population of clinical samples with a high probability, typically 95%. For example, whether the reference material is within (i) the ± 2SD range around the mean, (ii) within the 95% prediction interval around the regression line, or (iii) within the 95% confidence ellipse of multivariate statistics of the clinical samples [1].

Here, we describe a new approach that investigates whether the reference material is equivalent to the average clinical sample. It is based on the two-one-sided t-test approach (TOST, [2]) at a single target concentration, which requires that both reference material and clinical sample are concentration-matched. The TOST approach allows the calculation of a non-commutability budget under given conditions for analytical imprecision (5% α-error and 90% power).

For a minimum complexity protocol, in which each of 2 methods measures 2 clinical samples in singlicate (= group 1) and 1 reference material in duplicate (=group 2) commutability can be proven by

Table 1. Sample size calculations for non-commutability budgets using the TOST approach (conditions: analytical CV = 1.2%, α-error = 5%, power = 90%)

	Generic	**ALP**	**CHOL**	**CA**
Test/Bias limit[1]	5.39 x SD_{Δ}	6.7%	3%	0.82%
Ratio limit/SD_{Δ}	5.39	3.95	1.77	0.48
Budget (%)[2]	**n (per group)**			
0	2	3	8	94
10	3	3	9	99
20	3	3	10	117
30	3	4	12	151
40	3	4	16	205
50	4	6	23	295
60	5	8	35	460
70	8	13	62	816
80	16	29	138	1835
90	60	111	549	7337

[1] The limit in the first case is set at a minimum for proving commutability by TOST (when n = 2), i.e. 5.39 times SD_{Δ}, with SD_{Δ}= $SD_{\Delta CS}$= $SD_{\Delta RM}$ = $\sqrt{\frac{SD^2_{\Delta CS}+SD^2_{\Delta RM}}{2}}$ with CS = clinical samples, RM = reference materials.
For ALP, CHOL and CA the limit equals the (biological based) bias limit extracted from the Westgard Biodatabase.

[2] The non-commutability budget represents the % of the original test limit that can be consumed by non-commutability.

TOST, from setting the test limits to 5.39 times the standard deviation of the differences (SDΔ). Under these conditions no budget for non-commutability is allowed. This situation is the starting point in Table 1, and is referred to as test limit (5.39 x SDΔ) in the column with heading “generic”. As explained above, at this point, the non-commutability budget is 0%. Table 1 shows that with increasing sample size (calculated with the SAS Power and Sample Size software - SAS Institute Inc. 2007), increasing non-commutability budgets (expressed as a percentage of the test limit) can be tolerated. For example, with n = 16, 80% of the original test limit can be consumed by non-commutability. Table 1 further compares the generic test limit with 3 real-case limits reflected by the bias (%) that can be tolerated on the basis of the biological variation data listed in the Westgard Biodatabase, i.e., 6.7% for alkaline phosphatase (ALP), 3% for cholesterol (CHOL) and 0.82% for calcium (CA) [3]. Considering a typical analytical CV of 1.2% for the test methods (SDΔ = 1.697%), the ratio of the respective bias limits to the expected SDΔ amounts to 3.95% (ALP), 1.77% (CHOL) and 0.48% (CA). The table also shows the sample size required to allow increasing non-commutability budget (from 10% to 90%) for each of the cases. It is obvious that the lower the ratio, the higher the required sample/measurement size. For ALP with a ratio of 3.95%, which is close to the generic case, the sample sizes required for the tabulated non-commutability budgets are feasible for the single laboratory (n = 29 for a budget of 80%). In contrast, the feasibility of the required sample size 23 for the CHOL case is restricted to a 50% non-commutability budget, whilst for CA, already quite high sample sizes (94) are needed even without room for a non-commutability budget.

In conclusion, we showed that our t-testing approach can be used for commutability assessment with consideration of a desirable non-commutability budget in relation to the sample size. The test limits of the TOST could potentially be used as margins in equivalence tests [4] and depending on the ratio of the test limit versus the SDΔ, one must perform a low to high number of measurements to reach the desired budget. For quite low ratios this number becomes non-realistic for the single laboratory, but can be achieved by collective measurement of the reference material and clinical samples by several laboratories, as is possible in external quality assessment surveys [5].

References

1. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. Clin Biochem Rev 2007;28:139–47.
2. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. J Pharmacokin Biopharm 1987;15:657–80.
3. Westgard QC – Biodatabase. http://www.westgard.com/biodatabase1.htm (acessed July 2014).
4. Keller T, Weber S. Statistical test for equivalence in analysis of commutability experiments. Clin Chem Lab Med 2009;47:376–7.
5. Van Houcke S, Rustad P, Stepman H, Røraas T, Sandberg S, Thienpont L. Commutability assessment by use of external quality assessment surveys: a means to reduce the uncertainty in the commutability decision. Klinisk Biokemi I Norden 2013;25:32–7.

Broader international context, relevance and future perspectives

By providing a means to assess certain aspects of the quality of IVD assays, this thesis project has the potential to improve their quality. As part of the overarching Empower project, it can benefit three parties: (1) the individual clinical laboratory, (2) the IVD industry, and (3) the patient and healthcare authorities! In addition, this project creates a platform for improved communication between all involved parties.

1) The laboratory

The laboratory community can greatly benefit from the general information supplied by the Empower project. First of all the applications it provides can aid laboratories in achieving the quality required by ISO 15189. This international standard specifies requirements for the competence and quality of clinical laboratories. As such it is used in the development of quality management systems for the assessment of the laboratories' own quality and for use by accreditation bodies (for the endorsement and recognition of the competence of clinical laboratories, 1). Accreditation has, therefore, put a pressure on the laboratories to document their analytical quality. When a laboratory participates in the Empower project, it can fulfill some of the managerial and technical requirements set in the ISO standard. Indeed, the Percentiler and Flagger serve as a direct, real-time quality indicator for laboratory analyses of patient samples, as requested by the ISO 15189. By combining patient monitoring with their own IQC data, laboratories can create evidence about the mid- to long-term variation of their IVD instrument, calibrator, and reagent of the manufacturer (lot-to-lot consistency), and, thereby, provide the laboratory community with realistic quality specifications. This evidence is backed-up by information from other laboratories using the same assay. By collecting the information from a complete peer group, the laboratory's position is strengthened in claims versus the manufacturer. Furthermore, the Master Comparison studies give a calibration fix-point, information on the basic quality of assays and laboratory performance, as well as evidence on the standardization status and comparability of the examined assays. Collecting information from both the Percentiler and the Master Comparison surveys from several peer groups creates the possibility for cross-comparison between manufacturers, which allows insight into the quality and performance of other assays and may aid in selection of future test systems. Note, however, that the identity of the manufacturers is currently not being disclosed to all participants in the Percentiler/Flagger part of the Empower project. The laboratories receive information about the stability issues in their own peer group, but not of the other peers. This way of working was adopted in the hope to build a trust relationship with the manufacturers. When, however, stability issues remain to be long lasting problems the identity might still be disclosed.

In addition, laboratories which participate in the project can more transparently communicate with their clinicians. The Flagger, for example, has the potential to show the impact of performance instability on flagging rates. From this information, clinicians can better understand the fluctuations in flagging rate of results due to lot changes. It can also help the laboratory to keep the extent of the effect within desirable limits. Laboratories can also strengthen their position when clinicians attribute certain observations to an erroneous laboratory performance, e.g. increased diagnosis of certain diseases in the patient population they see. By consulting the Percentiler and Flagger data and observing the stability of performance in time, the laboratory is in a position to counter the clinician's assumption of erroneous laboratory performance.

2) The IVD manufacturer

IVD manufacturers can benefit from the Empower project in similar ways as the clinical laboratory. They can use the different aspects of the Empower project to assess the quality of performance of their IVD assays on real samples under "field" conditions. For the manufacturers, it is more relevant to discuss an analytical problem or effect of bias from patient results rather than from IQC data. The Percentiler provides this type of information and gives manufacturers additional insight on the mid- to long-term stability of their assays. In addition, it shows which assays are eligible for stability improvements. For example, when most assays under field conditions fail to meet stability specifications based on biological variation, this can be regarded as an incentive to improve the current state-of-the-art performance. Note, just like the laboratories, the manufacturers only receive information about the stability of their own assay (via peer group reports). The identity of other peer groups is not disclosed. Additionally, the Master Comparison surveys provide correct information on the comparability status of assays. Observations of non-comparability between manufacturers or assays can pinpoint the need for standardization efforts. Previously conducted surveys show that, even for "simple" clinical chemistry measurands, the standardization status of certain assays is still a matter of concern and there is much room for improvement (2).

Furthermore, it is a fact that the burden for a manufacturer is less when he needs to react on a group of laboratories presenting the same problem, rather than having to deal with a single laboratory having a particular issue. In this regard, compilation of the information from all project pillars and structuring according to peer group is utmost beneficial for the manufacturers.

3) The patient and healthcare authorities

Finally, also patients can benefit from the quality information the Empower project provides. As mentioned in the introduction, the individual patient will see the full test variation extending from manufacturer (including assay variants), over laboratories to system sites (including lot- and calibration variation). When quality of testing is poor, or when the introduced error is significant, it are ultimately the patients who suffer most. Therefore, proper patient monitoring and diagnosis requires sufficient between-assay comparability and sufficient stability of the used assays. Otherwise poor quality of IVD tests can, for example, lead to patients who are being misdiagnosed, who are treated for a longer time than necessary, who aren't treated when needed, etc. Also, poor quality of assays can lead to repetition of laboratory testing. This, in turn, leads to an unnecessary increase of healthcare costs and has a negative impact on the system's health economy (3).

"Big Data" projects like the Empower project have the potential of significantly improving assay stability and comparability and, thereby, lowering costs in healthcare. By showing the quality which is currently achieved by manufacturers and laboratories the project can critically inform healthcare authorities and the patients. Or to quote Sten Westgard on his essay on the Empower project (4):

"Empower IVD and other accuracy-based EQA/PT programs threaten to reveal significant method and quality differences, as well as critically inform laboratories. When the truth about method quality is revealed, some manufacturers will gain, but more manufacturers will lose; some laboratories will be pleased, but more laboratories will be unhappy to learn that they need to expend more effort and resources, change methods, change their operations, etc., in order to achieve the level of quality they thought they were already achieving."

References

1. International Organization for Standardization (ISO). ISO 15189: medical laboratories - requirements for quality and competence. 2012. http://www.iso.org/iso/catalogue_detail?csnumber=56115 (Accessed June 2016).
2. Horowitz GL. Assessing accuracy on the front lines: a pragmatic approach for single-donor proficiency testing. Clin Chem 2014;60:806-8.
3. Boone DJ. Is it safe to have a laboratory test. Accred Qual Assur 2004;10:5–9.
4. Westgard S. Is your lab empowered? 2015. https://www.westgard.com/lab-empowerment.htm (Accessed June 2016).

Summary and general conclusion

1. Summary

Chapter 1 – IQC monitoring

In the perspective of monitoring the stability of performance of IVD instruments, we discussed IQC for four analytes (calcium, sodium, inorganic phosphate, and albumin). We determined what type of information can be extracted from these stability studies. IQC data from Roche, which were evaluated daily, were visualized with Excel using a moving median grouped per 2, 5, 10, or 15 days. These data have proven to be ideal to focus on short-term (in)stability effects, however, they can also be used to address mid- to long-term effects. Daily IQC data also included the reagent container identification number of each measurement performed. With these data one can distinguish between analytical shifts which are caused by a new reagent container or by mathematical calibration (by plotting each measurement result in a scatterplot). Data provided by Bio-Rad only contained the monthly means of the QC materials which were measured daily. This type of data loses its ability for early recognition of measurement problems. Therefore, it can only be used to get a broad idea about the long-term performance of assays and/or IVD instruments.

Chapter 2 – The Percentiler

As a surplus to the established IQC in detecting analytical error, we discussed the Percentiler application. This freely available online web tool works with instrument-specific, daily outpatient medians. The latter are calculated and transmitted by participating laboratories from their middleware or LIS. The data are transmitted by e-mail, which is automatically read by our software and transferred into a MySQL database. In the pilot phase the project focused on 20 clinical chemistry analytes, FT4 and TSH, and on the instruments of five manufacturers (Abbott, Beckman, Ortho, Roche, and Siemens). Graphical and statistical interpretation of the data, allowed us to show that the Percentiler has the potential to visualize the analytical variation of measurement results from patient samples directly (instrument-specifically). It thereby provides real-time global evidence about mid- to long-term variation of the instrument, calibrator, and reagent. In addition, it can create evidence about the reasons for the observed variation: own performance (e.g. recalibration or instability) or manufacturer performance (e.g. lot-to-lot-variation). This evidence is backed-up by information from other laboratories using the same assay.

Furthermore, we showed that the Percentiler is a good tool to demonstrate what stability can be achieved by current state-of-the-art performance. The stability limits, covering the grey shaded zone in the Percentiler graphs, are guided by the systematic error limits based on biological variation. However, we took the current capability of diagnostic manufacturers into account and expanded these when necessary. The Percentiler observations showed that many laboratories are able to achieve long-term stability within these desirable bias limits. Test variability, however, due to calibration and lot changes is still a significant problem for several laboratories and manufacturers, especially for certain enzymes (ALT, AST). The data further confirms that the ion selective electrode technique for chloride (and to a lesser extent for sodium) still has robustness issues. In several private laboratories, pre-analytical problems jeopardize the reliability of potassium and LDH testing.

Finally, we proved that the Percentiler is useful for studying relative assay biases at median concentrations. This allows to assess the comparability of the different assays available on the market. Normally, the comparability status of the different assays can only be reliably assessed with an EQA study making use of commutable samples. However, by comparing the Percentiler peer group medians with the Master Comparison survey, we showed that the relative comparability between the peer groups was very similar in the Percentiler and Master Comparison data, for most analytes and peer groups.

Chapter 3 – The Flagger

Apart from the Percentiler, we discussed the Flagger application as an additional tool to monitor the (in)stability of assays. This tool makes use of the fact that the LIS alerts clinical biochemists about values which fall outside the reference interval, either lower or higher. The Flagger functionality is analogous to the Percentiler functionality. It is a freely available online web tool, which works with instrument-specific, daily outpatient hypo- and hyper-flagging rates. The latter are calculated and transmitted by laboratories from their middleware or LIS. The data are transmitted by e-mail, which is automatically read by software into a MySQL database. In the pilot phase the project focuses on the same analytes and peer groups as in the Percentiler. Participants can use the Flagger, just like the Percentiler, to monitor the stability of their instruments versus semi-arbitrarily chosen Flagger limits, which indicate what stability can be achieved at the lower and higher concentration ends. We also showed that combining the two online monitoring tools together has the additional potential of comparing the trends in patient medians with the trends in the hypo- and hyper flagging rate. Although a laboratory, thereby, still isn’t capable of translating the effect of limit violations on the clinical outcome, it can at least visualize the effect on the flagging rate.

Chapter 4 – The Master Comparison surveys

The Master Comparison surveys make use of freshly frozen single donation serum samples from 20 apparently healthy donors. These samples are as commutable as possible and are therefore ideally suited to address comparability between both laboratories within a single peer group and across assays. In addition, by measuring several analytes with a reference measurement procedure these studies are also able to demonstrate the standardization status of the different assays. Previously performed Master Comparison studies assessed the comparability status of albumin, calcium, creatinine, glucose, HDL-cholesterol, LDL-cholesterol, magnesium, phosphate, total protein, total cholesterol, triglycerides, and uric acid. In our 2014 study we focused our efforts on five common enzymes (ALP, ALT, AST, GGT and LDH) and three electrolytes (potassium, sodium and chloride) in eight different peer groups.

Our 2014 Master Comparison survey showed that peer performance and assay comparability were generally good for all electrolytes. An exception was the negative bias of the Roche assays for chloride. For enzyme assays both manufacturers and laboratories still have a lot to do to accomplish trueness of measurement. Peer performance strongly depended on the enzyme measured. ALT assays were particularly problematic, possibly due to the nature of the samples (concentrations typically in the reference interval). AMTM/RMP-corrected AMTM Sy/x values were similar to the peer estimates, indicating little influence of random sample-related effects. Comparability, or bias versus the RMP-corrected AMTM procedure, strongly depended on the enzyme measured. In general, the data show the need (i) to phase out the non-IFCC assay variants; (ii) to improve the quality, comparability, and standardization of the ALT assays; (iii) to improve the comparability of the GGT and LDH assays; (iv) to improve peer comparability at low concentration levels; (v) to aim for single-manufacturer efforts for improvement of comparability for nearly all enzymes studied.

Chapter 5 – The Empower project

To collect and communicate the combined information from the Master Comparison surveys, the Percentiler and the Flagger, the overarching Empower project was created. Its prrimary goal is to promote and improve the quality of global IVD testing by “bottom-up” cooperation of laboratories and IVD manufacturers. To accomplish these goals, the Empower project was founded on four pillars: (i) The Master Comparisons; (ii) virtual EQA 1: mid- to long-term monitoring of patient percentiles using the Percentiler and the Flagger; (iii) Virtual EQA 2: mid- to long-term monitoring based on IQC data; and (iv) conceptual and statistical education. Through the project we have been able to provide evidence on the intrinsic quality of assays and how this quality is sustained under field conditions (with data generated from real patient samples), to demonstrate how well assays and laboratories compare, and how stable they perform, and, in addition, to uncover all major bias components and sources. However, maybe most important is that the overarching Empower project functions as a communication and/or discussion platform of the Empower project team with both laboratories and IVD manufacturers.

Chapter 6 – Alternative statistical tools

The appropriate use of statistics is essential for correct data interpretation in the clinical laboratory. Several guidelines are available which aid clinical chemists for their local data processing and interpretation. These guidelines, however, have several limitations. The current C28 validation protocol for reference interval verification, for example, should be handled with caution. It tends to sacrifice power in order to avoid the error of falsely rejecting the reference interval. As an alternative we propose an optimized sampling protocol to avoid the chance of falsely accepting the existing reference interval. One possible approach could be to increase the sample size towards more statistical power.

Also the current EP14 commutability assessment protocol should be handled with caution, since methods with lower accuracy are less likely to detect a non-commutable reference material. As an alternative we propose a t-testing approach which can be used for commutability assessment with consideration of a desirable non-commutability budget in relation to the sample size.

2. General conclusion

Quality assurance in the clinical laboratory is essential to guarantee appropriate interpretation of clinical tests. In that regard, IQC and EQA have earned a well-established position. However, these tools don’t necessarily reflect the reality of patient testing due to non-commutability issues. We have, therefore, proven that the Empower project is a valid tool to circumvent that problem. When this project is integrated in the current QC environment, it can establish a bottom-up approach to quality assessment which facilitates collaboration between laboratories and IVD manufacturers. It has the potential to create reliable evidence about the intrinsic quality of assays and how this quality is sustained under field conditions.

The Percentiler and Flagger have proven to be efficient tools for real-time monitoring of the mid- to long-term stability of assays, instruments, calibrators, and reagents. In addition, they create evidence about the reasons for the observed variation (lab performance or manufacturer performance), and translate the effect of analytical instability on the surrogate medical outcome. Although some issues, such as population variation and stratification differences, must still be taken into account, clinical laboratories can use these tools to directly relate analytical instabilities to their effect on the patient medians and on the hypo- and hyper-flagging rates. Therefore, these tools can be applied as a stimulus for both laboratories and manufacturers to work towards improved assay stability, when needed.

The Master Comparison surveys, which focus on the current standardization and comparability status of commonly used commercial assays, showed that even for “simple” clinical chemistry measurands such as enzymes and electrolytes, the standardization and/or harmonization status of certain assays is still a matter of concern and there is much room for improvement.

From this perspective, we believe the Empower project is a new integrated tool for modern quality management which is of benefit to all stakeholders with interest in reliable laboratory data.

Samenvatting en algemene conclusie

1. Samenvatting

Hoofdstuk 1 – IQC monitoring

In het kader van stabiliteitsmonitoring van de prestatie van IVD instrumenten, hebben we IQC besproken voor vier analieten (calcium, natrium, anorganisch fosfaat, en albumine). We hebben bepaald welke informatie we kunnen extraheren op basis van deze stabiliteitsstudies. De IQC dataset van Roche, die dagelijks werd gemeten en geëvalueerd, werd gevisualiseerd met Excel aan de hand van een bewegende mediaan gegroepeerd per 2, 5, 10 of 15 dagen. Deze dataset was ideaal geschikt om korte-termijn (in)stabiliteitseffecten waar te nemen, maar de data was ook bruikbaar in het bestuderen van midden- tot lange-termijn effecten. De dagelijkse IQC dataset bevatte ook een identificatienummer van de container van de reagentia per meting. Met deze data kan men een onderscheid maken tussen de container- en kalibratie-effecten (door elk meetresultaat uit te zetten in een scatterplot). De IQC dataset van Bio-Rad rapporteerde enkel het maandelijks gemiddelde van de dagelijks gemeten QC materialen. Dit soort dataset verliest de mogelijkheid om analytische meetproblemen vroeg op te sporen en kan daarom enkel gebruikt worden om een idee te krijgen van de lange-termijn prestaties van assay's en/of IVD instrumenten.

Hoofdstuk 2 – De Percentiler

Bovenop de gevestigde IQC hebben we ook de Percentiler applicatie besproken om analytische fouten te detecteren. Deze gratis beschikbare online webtool werkt met dagelijkse instrument-specifieke medianen van ambulante patiënten. Die data wordt door een deelnemend labo berekend en verstuurd met hun middleware of LIS. De data wordt vervolgens verstuurd via e-mails, die automatisch ingelezen worden door onze software in een MySQL database. In de pilootfase van het project werd de focus gelegd op 20 analieten van de klinische chemie, FT4 en TSH en op de instrumenten van vijf fabrikanten (Abbott, Beckman, Ortho, Roche, en Siemens). Grafische en statistische interpretatie van de data toonde aan dat de Percentiler in staat is om de analytische variatie van meetresultaten van patiënten rechtstreeks te visualiseren (instrument-specifiek). Daardoor levert het wereldwijd bewijs over de midden- tot lange-termijn variatie van instrumenten, kalibratoren, en reagentia. Bovendien kan het bewijzen leveren over de reden van de geobserveerde variatie: kwaliteit van het labo (bv. herkalibraties of instabiliteit) of kwaliteit van de fabrikant (bv. variaties tussen verschillende loten). Dit bewijs wordt nog eens versterkt met informatie van labo's die dezelfde assay's gebruiken.

Daarnaast hebben we getoond dat de Percentiler een goeie tool is om te demonstreren welke stabiliteit bereikt kan worden met de huidige “state-of-the-art” prestaties. De stabiliteitslimieten, aangeduid met de grijze zone in de Percentiler grafieken zijn gebaseerd op de systemische fout limieten gesteund op het concept van biologische variatie. Maar, we hebben de huidige capaciteiten van de IVD toestellen in rekening gebracht en deze limieten verbreed wanneer dat nodig was. De Percentiler observaties toonden aan dat veel van de deelnemende labo’s in staat zijn om lange-termijn stabiliteit aan te houden binnen deze vooropgestelde limieten. Testvariabiliteit, als gevolg van kalibraties en lotwijzigingen, blijft helaas een significant probleem voor verschillende labo’s en fabrikanten. Zeker in het geval van bepaalde enzymen (ALT, AST). De data toont bovendien aan dat de ion-selectieve electrode techniek voor chloride (en in mindere mate voor natrium) nog steeds problemen heeft met de robuustheid. In sommige privélabo’s brengen preanalytiek problemen dan weer de betrouwbaarheid van kalium en LDH assay’s in gevaar.

Verder hebben we aangetoond dat de Percentiler kan gebruikt worden om de relatieve bias van de assay’s bij de mediaan concentratie te bestuderen. Dat maakt het mogelijk om de vergelijkbaarheid van de verschillende assay’s op de markt te beoordelen. Normaal gezien kan die vergelijkbaarheid van de assay’s enkel beoordeeld worden met een EQA studie die gebruik maakt van stalen die commutabel zijn. Maar, door de peer groep medianen van de Percentiler te vergelijken met de data van de Master Comparison studies, toonden we aan dat de relatieve vergelijkbaarheid tussen de peer groepen in de Percentiler sterk vergelijkbaar was met de Master Comparison data.

Hoofdstuk 3 – De Flagger

Naast de Percentiler, hebben we ook de Flagger besproken als extra tool om de stabiliteit van assay’s te monitoren. Deze tool maakt gebruik van het feit dat het LIS de klinisch bioloog alarmeert wanneer waarden buiten een referentie-interval vallen (zowel lager als hoger). De Flagger functioneert compleet analoog zoals de Percentiler. Het is een gratis beschikbare online webtool die werkt met dagelijks instrument-specifieke “hypo- en hyper-flagging rates”. Die data wordt door een labo berekend met hun middleware of LIS en wordt vervolgens verstuurd via e-mails, die automatisch ingelezen worden door onze software in een MySQL database. In de pilootfase werd de focus gelegd op dezelfde analieten en peer groepen als in de Percentiler. Deelnemers kunnen de Flagger dus op dezelfde manier gebruiken als de Percentiler om de stabiliteit van hun instrumenten te vergelijken met semi-arbitrair gekozen Flagger limieten, die aantonen welke stabiliteit kan bereikt worden bij de laagste en hoogste concentraties. Bovendien hebben we aangetoond dat, indien men de twee online monitoring tools combineert, men additioneel in staat is om de trends van de patiëntenmedianen te vergelijken met de trends van de “hypo- en hyper-flagging rates”. Alhoewel een labo dan nog steeds niet in staat is om het effect van limietoverschrijdingen te vertalen in het klinische effect, kan het ten minste het effect op de “flagging rate” visualiseren.

Hoofdstuk 4 – De Master Comparison studie

Deze Master Comparison studies maken gebruik van vers ingevroren enkel-donatie serumstalen van 20 schijnbaar gezonde individuen. Deze stalen zijn zo commutabel als mogelijk en zijn daarom ideaal geschikt om de vergelijkbaarheid van labo’s binnen en tussen peer groepen te onderzoeken. Bovendien, doordat sommige analieten werden gemeten met een referentieprocedure, is het ook mogelijk om deze studies te gebruiken om de standaardisatiestatus van verschillende assay’s te demonstreren. Master Comparison studies, die voorafgaand aan dit doctoraal werk werden uitgevoerd, bestudeerden de vergelijkbaarheidsstatus van albumine, calcium, creatinine, glucose, HDL- en LDL-cholesterol, magnesium, fosfaat, totaal-eiwit, totaal-cholesterol, triglycriden en urinezuur. In de recente studie uitgevoerd in 2014 legden we de focus op vijf enzymen (ALP, ALT, AST, GGT, en LDH) en drie elektrolyten (kalium, natrium en chloride) in acht verschillende peer groepen.

Onze 2014 master Comparison studie toonde aan dat peer prestaties en assay vergelijkbaarheid doorgaans goed zijn voor alle elektrolyten. Een uitzondering was de negatieve bias van de Roche assay’s voor chloride. In het geval van enzymen hebben zowel fabrikanten als labo’s nog veel te verwezenlijken met betrekking tot de “trueness” van hun metingen. De prestaties van de peer groepen waren sterk afhankelijk van het gemeten enzym. Vooral ALT assay’s waren problematisch, vermoedelijk door de aard van de stalen (concentraties binnen het referentie-interval). AMTM en RMP-gecorrigeerde AMTM Sy/x waarden waren vergelijkbaar aan de geschatte peer waarden, wat aanduidt dat er weinig invloed optrad van staal-gerelateerde effecten. De kwaliteit van de vergelijkbaarheid, of bias ten opzichte van de RMP-gecorrigeerde AMTM procedure, was sterk afhankelijk van het gemeten enzym. In het algemeen toonden de data nood aan (i) het verwijderen van niet-IFCC assay varianten; (ii) het verbeteren van de kwaliteit, vergelijkbaarheid en standaardisatie van ALT assay’s; (iii) het verbeteren van de vergelijkbaarheid van de GGT en LDG assay’s; (iv) het verbeteren van de peer vergelijkbaarheid bij lage concentraties; (v) het streven van de fabrikanten om de vergelijkbaarheid van enzymen te verbeteren.

Hoofdstuk 5 – Het Empower project

Om alle informatie van de Master Comparison studies en van de Percentiler en Flagger samen te bundelen en te verspreiden werd het Empower project gecreëerd. Het doel van dit project is om de vergelijkbaarheid en stabiliteit van labotesten te bestuderen en te verbeteren via bottom-up samenwerking tussen labo’s en IVD fabrikanten. Om deze doelen te vervullen, werd het project onderverdeeld in vier onderdelen: (i) de Master Comparisons; (ii) virtueel EQA 1: midden- tot lange-termijn monitoring van patiëntenpercentielen met behulp van de Percentiler en Flagger; (iii) Virtueel EQA 2: midden- tot lange-termijn monitoring gebaseerd op IQC data; en (iv) conceptuele en statistische educatie. Met dit project zijn we in staat om bewijs te leveren over de huidige intrinsieke kwaliteit van assay’s en of die kwaliteit stand houdt onder labocondities (met data gegenereerd van echte patiëntenstalen). Met dit project zijn we in staat te demonstreren hoe goed assay’s en labo’s vergelijken met elkaar, en hoe stabiel ze presteren, en zijn we instaat om alle voorname bias componenten op te sporen. Maar bovenal functioneert het Empower project als een communicatie-, en discussieplatform tussen het Empower project team en zowel de labo’s als de IVD fabrikanten.

Hoofdstuk 6 – Alternatieve statistische tools

Het gebruik van statistiek is essentieel voor correcte interpretatie van de data in een klinisch labo. Verschillende gebruiksaanwijzingen zijn beschikbaar die de klinisch bioloog helpen om hun lokale data te verwerken en interpreteren. Deze gebruiksaanwijzingen hebben helaas een aantal limieten. Het huidige C28-validatieprotocol om referentie-intervallen te interpreteren, bijvoorbeeld, moet voorzichtig gehanteerd worden. Dit protocol prefereert het opofferen van statistische power om te vermijden dat men het correcte referentie-interval verkeerdelijk verwerpt. Als alternatief stellen wij een geoptimaliseerd staalnameprotocol voor om te vermijden dat een verkeerd referentie-interval verkeerdelijk wordt aanvaard. Een mogelijke aanpak is onder andere het verhogen van het aantal stalen om meer power te krijgen.

Ook het huidige EP14 protocol om commutabiliteit te verifiëren moet voorzichtig gehanteerd worden. Methodes met een slechte accuraatheid hebben minder kans om een niet-commutabel referentiemateriaal te detecteren. Als alternatief stellen we een aanpak voor die gebruik maakt van een soort t-test. Deze aanpak verifieert de commutabiliteit rekening houdend met een zeker niet-commutabiliteitsbudget in relatie tot de grootte van de staalname.

2. Algemene conclusie

Het garanderen van de kwaliteit in een klinisch labo is essentieel om correcte interpretatie van klinische tests te verzekeren. In dat opzicht hebben interne- en externe kwaliteitscontrole een belangrijke rol verworven in het klinisch labo. Helaas kunnen deze tools niet altijd de realiteit van de patiëntendata visualiseren omwille van commutabiliteitseffecten. Daarom, hebben we aangetoond dat het Empower project een geldig middel is om dat probleem te omzeilen. Wanneer dit project wordt geïntegreerd in huidige kwaliteitscontrole-omgevingen, dan kan het een "bottom-up" aanpak voor kwaliteitscontrole garanderen, die een samenwerking tussen labo's en fabrikanten bevordert. Dit project heeft het potentieel om betrouwbaar bewijs te creëren over de intrinsieke kwaliteit van assay's en hoe die kwaliteit wordt behouden in labocondities.

De Percentiler en Flagger zijn efficiënte middelen om op real-time basis de midden- tot lange-termijn stabiliteit van assay's, instrumenten, kalibratoren en reagentia op te volgen. Bovendien creëren ze bewijs over de redenen van de geobserveerde variatie (prestaties van het labo of van de fabrikant) en vertalen ze het effect van analytische instabiliteit op surrogaat medische uitkomsten. Ondanks de aanwezigheid van enkele problemen, zoals hoge populatievariatie en verschillen in stratificatie, kunnen klinische labo's deze tools gebruiken om het effect van analytische instabiliteit te observeren op patiëntenmedianen en "hypo- en hyper-flagging rates". Daarom kunnen deze applicaties aangewend worden als stimulans voor zowel labo's als fabrikanten om de stabiliteit van assay's te verbeteren (indien nodig).

De Master Comparison studies, die focussen op de standaardisatie- en vergelijkbaarheidsstatus van commerciële assay's, toonden dat er zelf voor eenvoudige klinische chemie analieten, zoals enzymes en elektrolyten nog veel ruimte voor verbetering is, met betrekking tot de standaardisatie- en/of harmonisatiestatus.

Vanuit dit perspectief geloven we dat het Empower project een welkome nieuwe tool is voor integratie in het moderne kwaliteitssysteem ten voordele van alle belanghebbende partijen met interesse in betrouwbare laboratoriumdata.

Curriculum Vitae

Personal information

Name	**Goossens Kenneth**
Birth	26 December 1989, Lokeren
Nationality	Belgian
Address	Groendreef 24 Bus 3 B-9160 Lokeren

Studies

2001-2005	Secondary school – Latin – Sint-Teresia, Lokeren
2005-2007	Secondary school – Math-science – Sint-Lodewijks, Lokeren
2007-2012	Master biochemistry and biotechnology– University Ghent

Experience

01/10/2011 – 30/05/2012	Master thesis – PSB, VIB Gent
01/08/2013 – present	Assistant – Laboratory for Analytical Chemistry, University Ghent

Scientifical publications

Published	9
Submitted	1

Lectures

Concerning a European project	6
At a national scientific manifestation	0
At an exchange project	0

Poster presentations

At an international scientific manifestation	4

Participation at a congress, scientific manifestation

International	5
National	0

1. Research

1.1 Title PhD thesis:

Modern tools for improving comparability and stability of *in vitro* diagnostic tests

1.2 List of A1-publications

1. Goossens K, De Grande L, Keller T, Weber S, Thienpont L. Verification of reference intervals by the C28 protocol – The alpha error/power trade-off. Clin Chim Acta 2014;436:18–19

2. Goossens K, Van Uytfanghe K, Thienpont L. Calculation of non-commutability budgets by t-testing. Clin Chim Acta 2015;438:212–213

3. Goossens K, Thienpont L. Reference measurement procedure corrected all method trimmed mean - The best of two worlds. Clin Chim Acta 2015;440:55–56

4. Goossens K, Brinkmann T, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Clin Chem Lab Med 2015;53(10):e269-70

5. De Grande L, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont L. The Empower Project – A new way of assessing and monitoring test comparability and stability. Clin Chem Lab Med 2015;53(8):1197-204

6. Goossens K, Van Uytfanghe K, Thienpont L. Trueness and comparability assessment of widely used assays for 5 common enzymes and 3 electrolytes. Clin Chim Acta 2015;442:44-5

7. Goossens K, Van Uytfanghe K, Twomey P, Thienpont L, and Participating laboratories. Monitoring laboratory data across manufacturers and laboratories – A prerequisite to make "Big Data" work. Clin Chim Acta 2015;445:12-18

8. De Grande L, Goossens K, Van Uytfanghe K, Das B, MacKenzie F, Patru M, Thienpont LM, for the IFCC Committee for Standardization of Thyroid Function Tests (C-STFT). Monitoring the stability of the standardization status of FT4 and TSH assays by use of daily outpatient medians and flagging frequencies. Clin Chim Acta 2016; doi: 10.1016/j.cca.2016.04.032. [Epub ahead of print]

9. De Grande L, Goossens K, Van Uytfanghe K, Halsall I, Yoshimura NJ, Hens K, Thienpont LM. Using "Big Data" to describe the effect of seasonal variation in thyroid-stimulating hormone. Clin Chem Lab Med 2016; 55(2):e34-e36

1.3 Scientific lectures

1.3.1 Lectures

1. Goossens K. New trends in EQA – Master comparisons with native sera. IFCC Worldlab – 22nd international congress of clinical chemistry and laboratory medicine (22-26 June 2014), Istanbul, Turkey.

2. Goossens K. New trends in EQA – Master comparisons with native sera. EuroLabFocus – 3rd EFLM-UEMS congress (7-10 October 2014), Liverpool, UK.

3. Goossens K. Master Comparison 2014 and Future. Empower IVD•GLOBE Master Comparisons and Patient Percentile Monitoring workshop (10 December 2014), Ghent, Belgium.

4. Goossens K. The Percentiler and Flagger – Mid- to long-term quality monitoring. GLIMS User Meeting (3 June 2015), Utrecht, The Netherlands.

5. Goossens K. Current status of the Empower project & The Percentiler and The Flagger: Mid- to long-term quality monitoring. Empower IVD•GLOBE Status and future direction workshop (9 December 2015), Ghent, Belgium.

6. Goossens K. The Percentiler and the Flagger application – New tools for mid- to long term quality management. MIPS customer event (18 February 2016), Ghent, Belgium.

1.3.2 Posters

1. Goossens K, Stöckl D, Van Uytfanghe K, Thienpont L. Hierarchy of target values and acceptance limits in EQA. IFCC Worldlab – 22nd international congress of clinical chemistry and laboratory medicine (22-26 June 2014), Istanbul, Turkey.

2. Goossens K, Stöckl D, Van Uytfanghe K, Thienpont L. Hierarchy of target values and acceptance limits in EQA. EuroLabFocus – 3rd EFLM-UEMS congress (7-10 October 2014), Liverpool, UK.

3. Goossens K, De Grande L, Stöckl D, Van Uytfanghe K, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. Euromedlab – 21st IFCC – EFML European congress of clinical chemistry and laboratory medicine (22-24 June 2014), Paris, France.

4. Goossens K, De Grande L, Stöckl D, Van Uytfanghe K, Thienpont L. On-line flagging monitoring – A new quality management tool for the analytical phase. 2015 AACC Annual Meeting and Clinical Lab Expo (26-30 July 2015), Atlanta, US.

2. Educational experience

2.1 Practical exercises

- Academic year 2013-2014/2014-2015/2015-2016
  Instrumental Analytical Chemistry (2nd/3rd bachelor): 6 weeks – 25h/week (preparation and support)

3. Administrative and scientific services

- Supporting scientific services for the ID-MS reference lab

Acknowledgements

As most people know, I'm not very good with choosing the right words. Instead of looking for aspiring words of gratitude in other acknowledgements, I will give the most sincere thank you I can give in my own way.

THANK YOU

TO

FAMILY

Parents:

Frankie Goossens & Christel Vermeulen

Brother:

Cedric Goossens

Grandparents:

Wilbert Goossens & Lea Vanpeteghem

MANAGEMENT AND COWORKERS

Promotor & adviser

Prof. Dr. Linda Thienpont &

Dr. Dietmar Stöckl

Coworkers

Dr. Katleen Van Uytfanghe &

Linde De Grande

PARTICIPANTS TO THE EMPOWER PROJECT WORLDWIDE