

Why Metrological Traceability Matters in Medical Laboratory Diagnostics

Marith van Schrojenstein Lantman ^a, Christa Cobbaert ^b, Mauro Panteghini ^c,
Miranda van Berkel ^{a,d}, Ruben L. Smeets ^d, Jaap J. van Hellemond ^{a,e}
and Marc H.M. Thelen ^{a,d,*}

Application of the results provided by medical laboratories plays an essential role in medical decision-making. This is not limited to diagnosis and monitoring of disease but also involves its use in other phases of the health continuum, e.g., predisposition, risk stratification, screening, staging, prognosis, and surveillance. With the growing importance of precision medicine, the importance of requirements related to clinical performance, and consequently analytical performance of laboratory tests, also grows. To allow the community of laboratory medicine to translate clinical need into a test arsenal with adequate performance, the application of metrology concepts is essential. This paper summarizes, for all steps in the examination process from test development to clinical interpretation, why and how metrological traceability is a fundamental requirement for adequate medical decision-making and is critical for correct use of test results in algorithms and artificial intelligence-led approaches. This includes the importance of metrology concepts and their correct implementation for obtaining equivalence of test results upon cross-facility result exchange for primary or secondary use in healthcare and research. This is not limited to biochemistry and hematology but is also of importance to other areas of laboratory medicine, including microbiology. This paper provides an overview of the purposes of the underappreciated science of metrology in modern laboratory medicine and its importance to patients and caregivers.

INTRODUCTION

Over the last decades, “clinical chemistry” has gradually developed into laboratory medicine. This transition had its origin in 2 important trends. The first is that the development of medical

laboratory measurement procedures increasingly relied on the in vitro diagnostics (IVD) industry producing CE-marked (Conformité Européenne) and/or Food and Drug Administration (FDA)-approved IVD products, rather than on laboratory specialists developing laboratory-developed tests (LDTs). The

^aSKML, Foundation for Quality Assessment in Medical Laboratory Diagnostics, Nijmegen, the Netherlands; ^bDepartment of Clinical Chemistry and Laboratory Medicine, Leiden University Medical Center, Leiden, the Netherlands; ^cDepartment of Laboratory Medicine, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, Torun, Poland; ^dDepartment of Laboratory Medicine, Radboud University Medical Center, Nijmegen, the Netherlands; ^eDepartment of Medical Microbiology, Erasmus University Medical Center, Rotterdam, the Netherlands.

*Address correspondence to this author at: SKML, Radboud University, Toernooiveld 300, Nijmegen 6535EC, the Netherlands. E-mail mthelen@skml.nl.

Disclaimer: An opinion paper by the organizers and speakers of a symposium on the why, how, and what of metrology as the fundamental requirement for equivalent test interpretation in laboratory medicine, organized by the Dutch foundation for external quality assessment (SKML).

Received October 8, 2025; accepted November 17, 2025.

<https://doi.org/10.1093/jalm/jfaf203>

© The Author(s) 2026. Published by Oxford University Press on behalf of Association for Diagnostics & Laboratory Medicine.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

second trend is that the scientific societies for clinical chemistry gradually became aware that the value of their profession would benefit from more interaction of their members with requesting clinicians in relation to the pre- and post-analytical phases. The training of clinical chemists therefore shifted from a focus purely on analytical chemistry toward an increasing role in medical consultation. As a consequence, the analytical phase and its impact on patient care has tended to receive less focus, because it has been considered, not always correctly, that the analytical phase was prone to a smaller number of errors (e.g., errors caused by non-harmonized results) than other parts of the testing cycle (1,2).

However, with the continued evolution of laboratory medicine, the role of laboratory professionals in test development is receiving renewed attention. Over the last decade, test development has increasingly relied on selective methods requiring specific definition of the measurand (the quality intended to be measured) and its distinction from possible interfering cross-reactants. Another change is related to the use of laboratory results outside the boundaries of the organization in which the results were produced. Both these developments have stimulated the application of several aspects of metrology to safeguard correct and equivalent interpretation of laboratory results as a correct answer to a specific clinical question (3).

In the past, people could argue that metrological traceability was not always feasible in laboratory medicine due to a lack of clear indications about how to proceed with its implementation. However, the release of the 2020 revision of ISO17511 on standardization in laboratory medicine has taken away such cause for objection. That standard now provides scenarios for all different combinations of available components in the metrological traceability chain, including the contribution of all steps involved to the uncertainty of measurement (4). Based on that

document, both practical and educational literature is available for its correct implementation (3,5,6).

This collective opinion paper summarizes the lectures of a symposium that was organized to illustrate the importance of metrology for several aspects in laboratory medicine. The symposium, organized as the introduction to the PhD defense of the first author on the relationship between analytical performance and clinical decision-making (7), aimed to create awareness for the need to (re)discover metrological traceability as a central and essential part of the knowledge domain of modern laboratory medicine.

METROLOGICAL TRACEABILITY OF MEDICAL TEST RESULTS STARTS WITH DEFINITION OF THE MEASURAND

Knowing the measurands, i.e., the specific quantities or properties to be measured in medical tests, is fundamental to ensuring accurate, meaningful, and actionable healthcare decisions. Measurands define not only the intended analyte and quantity but also the phenomenon, body, or substance that carries it, such as the sodium concentration in blood, systolic pressure in arteries, or viral load in a sample (4). Without clarity on the measurand, even precise numerical results can be misleading or misinterpreted. In clinical practice, measurands guide diagnosis, treatment planning, and monitoring. For instance, measuring blood pressure without specifying whether it is systolic or diastolic, or under what conditions it was taken, can lead to incorrect conclusions and inappropriate interventions. Similarly, laboratory tests must clearly define the analyte, matrix, and the selectivity of the method used to ensure consistency and comparability across laboratories. Moreover, understanding measurands is essential for metrological traceability, which links test results to standardized references and ensures reliability over time and across settings. Metrological traceability supports

regulatory compliance, patient safety, and scientific integrity. Insufficiently defined or ambiguous measurands can result in misdiagnoses, ineffective treatments, and increased healthcare costs due to unnecessary procedures or repeat testing. Ultimately, precise knowledge of measurands empowers clinicians to interpret results correctly, fosters transparency in patient communication, and enhances trust in medical testing systems.

DEFINING THE MEASURANDS IN A CHANGING LANDSCAPE

In the 21st century, citizens and patients live in a global world and are confronted with ongoing evolution in science, metrology, and technology. Alignment between these three domains is key in laboratory diagnostics and implies that the definition of the measurand may be a moving target. Over the last 3 to 4 decades, approaches to analysis have evolved from measuring metabolites, substrates, and enzymes with *in-house* tests based on chemical method principles into more selective, automated (first batch-wise and later random access) methods and more recently into highly selective mass spectrometry-based methods. Moreover, this era of precision medicine also reveals an unmet clinical need for molecular definitions of health and disease, with unequivocally defined measurands (8). Ultimately, the rapid evolution of data science, with integration of conventional risk factors and emerging risk-enhancing factors, brings along improved risk stratification in, for example, primary prevention for cardiovascular disease (9) with reclassification of patients for individualized management and improved patient outcomes. In addition, the gradually increasing number of regulations affecting healthcare policies brings along new requirements, such as those for secondary use of medical test results in the European Health Data Space (EHDS) regulation and the European Union (EU) Artificial Intelligence Act.

WHY DEFINING MEASURANDS MATTERS

So far, medical laboratories in hospitals mostly rely on single biomarkers in medical tests. Consequently, there are high expectations about the analytical and clinical performance of these “uniplex” medical tests, especially those having a decisive role in clinical care pathways. Examples are serum creatinine and urine albumin for kidney function; cardiac troponins, C-reactive protein, and natriuretic peptides for heart disease; liver enzymes and coagulation factors for liver disease; hemoglobin (Hb) A_{1c} for diagnosis and monitoring of diabetes mellitus.

As intended uses of tests stated in clinical guidelines and IVD information for use evolve, it is essential that laboratory specialists periodically re-assess the measurand definition and its associated measurement uncertainty (MU). This is also needed when changing methods, reagents, and/or instruments which have different selectivity and performance characteristics. Below we present some representative examples, which illustrate the need for molecularly defined measurands in relation to their intended role in the medical process and the need for the selectivity to distinguish clinically relevant measurands from clinically irrelevant measurands or interferences.

Example 1: Low-Density Lipoprotein Cholesterol Defined with an Operational Definition Based on Ultracentrifugation

Low-density lipoprotein cholesterol (LDL-c) is considered a surrogate for the LDL load. LDL is defined based on beta-quantification ultracentrifugation encompassing a continuum of lipoproteins with a density ranging from 1.019 to 1.063 g/mL. The genetically determined lipoprotein(a) (Lp(a)) has partially overlapping density (1.05 to 1.12 g/mL). Cholesterol is measured in LDL, which is defined as LDL-c. LDL-c standardization is overseen by the Cholesterol Reference Method Laboratory Network run by the U.S. Centers for Disease

Control and Prevention. This definition of the measurand was successful when LDL-c concentrations in the populations were high and were treated with, for example, statins. The National Education Cholesterol Program (NCEP) performance criteria were defined for LDL-c in the context of treatment goals. With the global rise of cardiovascular disease (CVD) and with the successful introduction of more powerful cholesterol-lowering medications, such as proprotein convertase subtilisin/kexin type 9 (PCSK9)-inhibitors, very low LDL-c levels on treatment (approximately 1 mmol/L) became the treatment goal. At these low LDL-c levels, irrespective of direct or calculation-based methods, MU increases and traditional LDL-c measurements are no longer fit for purpose and should be replaced by a defined alternative (10). Sticking to LDL-c also impacts patient care. Quoting Jim Otvos: “a problem in need of recognition, acknowledgement and, ultimately, a solution, is the continued reliance on LDL-c as the standard measure of LDL and LDL-related cardiovascular risk. The problem in a nutshell is that LDL-c is a seriously flawed measure of LDL, and the extent to which many risk markers contribute to CVD risk “beyond LDL” is distorted by the traditional use of LDL-c as the LDL representative in multivariable risk prediction models.” (11).

The European Atherosclerosis Society (EAS) and the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) joint consensus report on quantifying atherogenic lipoproteins (12) has recommended apolipoprotein B (apo B) as a secondary treatment target in patients with mild to moderate hypertriglyceridemia, in whom LDL-c measurement or calculation is inaccurate and often less predictive of cardiovascular risk. The authors’ reasons for being hesitant about apoB as a target are the lack of evidence on cost-effectiveness of apo B-guided treatment. A former American Association for Clinical Chemistry (AACC) Lipoproteins and Vascular Diseases Division Working Group on Best Practices position statement on apo B reviewed the evidence for this

but similarly stopped short of recommending replacement of LDL-c as the primary target (13). The authors’ reason was pragmatic and understandable, a reflection of the “inconvenient” part of this particular truth: “changing perceptions and practices will not be easy.” So, they suggested that, for the time being, apo B should be used “along with LDL-c.”

Defining measurands is especially challenging in the case of heterogeneous measurands. A typical example is the size polymorphism of the unique measurand apolipoprotein A (apo(a)) in Lp(a). Polyclonal Lp(a) immunoassays that partially cross-react with the KIV₂ repeats in apo(a) masked the association between Lp(a) levels and CVD in the nineties, which led to false-negative conclusions about its clinical utility (14). Moreover, expressing Lp(a) in mass units instead of molar units is a flawed concept, giving Lp(a) the reputation of a massively misunderstood metric. Yet, with the introduction of mass spectrometry in diagnostic laboratories and the selection of defined proteotypic apo(a) measurands out of the KIV₂ repeat range, accurate molar apo(a) results may be generated that reveal a continuous association with CVD.

Example 2: Immunosuppressive Drug Monitoring Results Confounded by Flawed Immunoassays

In the organ transplantation domain, immunosuppressive drug monitoring (ISD) is key to preserving transplanted organ function during a person’s lifetime. Many case reports are available in the scientific literature regarding the shortcomings of current ISD immunoassays, with devastating effects on patient management and outcome (15). Consequently, defining ISD measurands and measuring them selectively matters to ensure accurate dose adjustments; minimize toxicity and rejection risk; enable consistent interpretation across laboratories and time; and support traceability and quality assurance.

Apart from these examples, many other situations exist where measurand definition and selectivity for the intended measurand make the difference

between undefined numbers and relevant results. In hemostasis testing for instance, antithrombin (AT) tests using selective mass spectrometry have added value when used in a complementary manner to a conventional AT-activity test for detecting AT deficiency. It has become obvious that detection and quantification of distinct molecular forms may unravel clinically relevant AT mutants (e.g., causing recurrent pregnancy loss) (16).

BEST PRACTICES FOR DEFINING THE MEASURAND BASED ON THE STATE OF SCIENCE, TECHNOLOGY, AND METROLOGY

Defining measurands accurately in medical testing is essential for ensuring reliable, interpretable, and clinically meaningful results. A measurand definition checklist for laboratory professionals is presented in [Table 1](#), as a tool for making rational choices about the measurands to be measured.

ANALYTICAL PERFORMANCE SPECIFICATIONS FOR LABORATORY MEASUREMENTS BASED ON MEDICAL REQUIREMENTS TO ESTIMATE HOW GOOD IS GOOD ENOUGH

Obtaining harmonization of laboratory results is an absolute priority for public health. The goal is to obtain “accurate measurement,” which is an unbiased measurement associated with a suitable MU related to a recognized standard, as described in ISO 17511:2020 (4). To define the suitability of MU for the clinical application of the measurement, analytical performance specifications (APS) must be predefined as a set of criteria that specify the quality required for values assigned to a clinical sample to satisfy clinical needs (17). Essential concepts for correct APS definition were described during the EFLM Strategic Conference held in

Milan in 2014. In particular, 3 models to set APS were proposed, with the preference for model selection primarily directed toward the measurand and its biological and clinical characteristics (18).

The first model, based on the effect of analytical performance on clinical outcomes, is applicable to measurands with a well-defined role in the diagnosis of a specific disease. However, directly connecting laboratory testing to patient outcomes is challenging, so that indirect approaches considering the impact of analytical performance of the test on clinical (mis)classifications or decisions and thereby on probability of outcomes, using simulation or decision analysis, have been proposed (19). Cardiac troponin testing can be used to exemplify measurands that should be allocated to the outcome-based model for deriving APS, as they have a central role in decision-making regarding acute coronary syndrome. The influence of assay variability on the number of false-positive and false-negative results for acute myocardial infarction diagnosis was evaluated using simulation models (20). From this information, IVD stakeholders can derive APS for troponin measurements. Other examples of APS derived using indirect outcome-based models are available for measurands to be allocated to this model (3). In the absence of information about outcome-based APS for measurands that theoretically should be allocated to this model, temporary allocations to one of the other two models described below should be considered, according to the measurand characteristics (18).

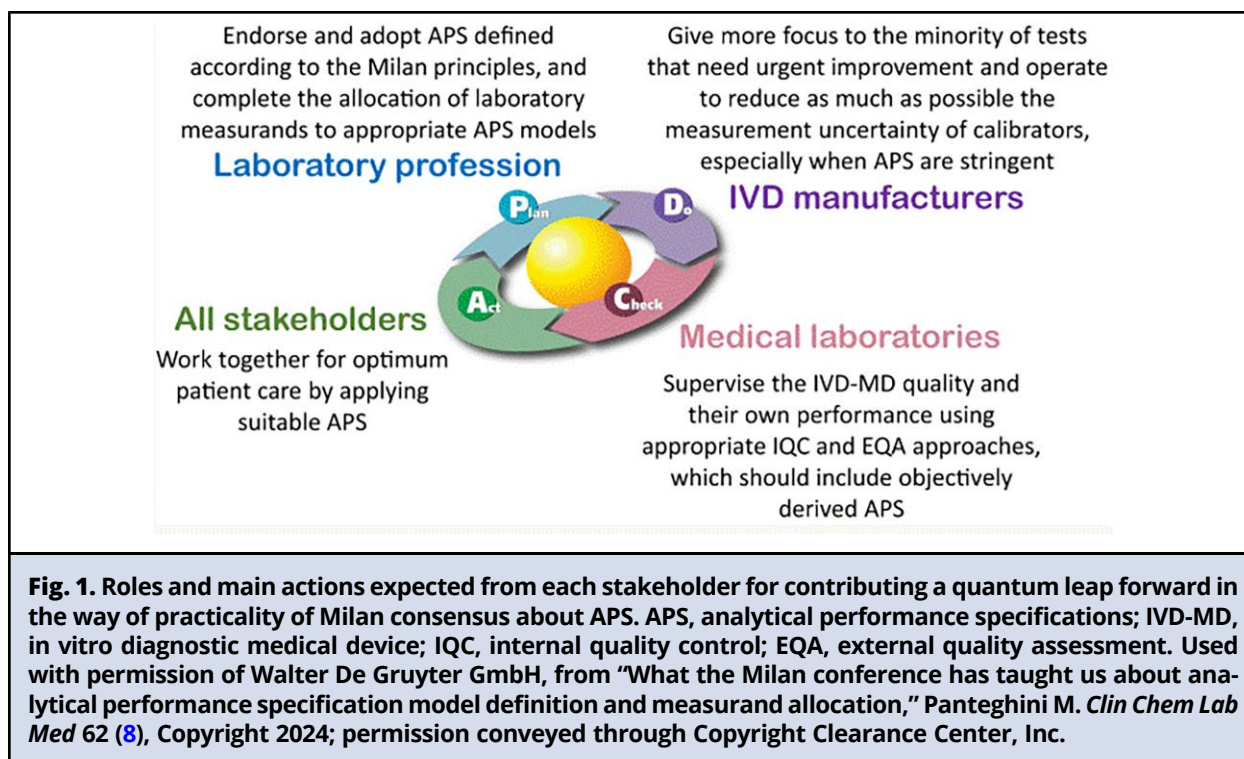
The second model is based on the biological variation (BV) of the measurand and should be applied to measurands with high homeostatic control (e.g., plasma ions) or if a measurand has stable concentrations when a person is in good health (e.g., serum creatinine). In contrast, it should not be used for measurands without steady-state status such as, for instance, some hormones and urine analytes. The BV-based APS model aims to minimize analytical noise relative to the BV of the measurand by defining APS as some fraction of the BV. The EFLM Working Group on BV has generated a database

Table 1. Measurand definition checklist for medical laboratories.	
Aspect	Checks to be performed
1. Clear identification	Is the measurand explicitly named (e.g., “glucose concentration”)? Is the biological matrix specified (e.g., plasma, serum, urine)? Is the measurement unit defined (e.g., mmol/L, g/L)? Is the method of measurement stated (e.g., enzymatic assay, immunoassay)?
2. Metrological traceability	Is the measurand traceable to a reference measurement system (e.g., certified reference material, internationally agreed reference measurement procedure)? Are calibration procedures documented and aligned with international standards (e.g., ISO 17511:2020)?
3. Measurement conditions	Are environmental and procedural conditions specified (e.g., temperature, pH, timing)? Are pre-analytical variables (e.g., fasting status, sample handling) controlled and documented?
4. Measurement uncertainty estimation	Is measurement uncertainty evaluated and reported? Is the method for uncertainty estimation (e.g., bottom-up or top-down) documented?
5. Biological variation consideration	Are biological variation data usable to set performance specifications? Are reference intervals and decision limits based on population-specific data?
6. Standardized terminology	Are definitions consistent with the International Vocabulary of Metrology (VIM)? Are terms aligned with ISO 15189:2022 and CLSI guidelines?
7. Validation and verification	Has the measurand been validated for scientific validity, clinical and analytical performance? Are periodic reviews conducted when methods, reagents, or instruments change?
8. Documentation and communication	Is the measurand definition included in laboratory standard operating procedures and test reports? Is the measurand definition clearly communicated to clinicians and other laboratory end users?
9. Quality control integration	Are internal quality control materials aligned with the measurand definition? Are external quality assessment schemes used to verify measurand accuracy?

Table 2. Remaining limitations of the EFLM biological variation database.
Impact of differences in selectivity of methods used for some measurands in different studies is not considered.
Influence of insufficient method sensitivity to detect measurands present in plasma at very low concentrations in all samples of all subjects enrolled in the BV protocols is not considered.
Although studies are weighed according to the quality of the study-design, the employed strategy based on the meta-analysis of available data may still lead to flaws, due to inclusion of studies with significant heterogeneity and low quality.
Lacking instruction for which measurands APS based on BV should not be used and other models are therefore preferable.

with essential information about the BV and derived APS for different measurands. However, some limitations remain and should be carefully considered using the listed information (Table 2) (3).

The third model, based on the state of the art of the measurement, should be applied when a measurand has neither a central diagnostic role nor strict homeostatic control. Furthermore, this



model can be temporarily used for those measurands still waiting for the definition of outcome-based APS or for which the BV-based model should not be used because strict homeostatic control is lacking. The state-of-the-art-based model applies for most urinary measurands, such as sodium, potassium, calcium, magnesium, etc., for which the concentrations may vary widely to maintain the corresponding plasma concentrations stable. Drawbacks for the state-of-the-art concept have been highlighted, in particular in relation to the difficulty in defining the level of analytical performance to be considered as state of the art, the "instability" of the information describing analytical performance (which can change for the better but also deteriorate), and the lack of direct relationship between what is analytically achievable and what is clinically needed (21).

Finally, drugs represent a special category of measurands that need a dedicated approach for

deriving APS, based on fundamental pharmacokinetic theory and elimination half-life of the drug. Based on these concepts, a specific model has been proposed (22).

Recently, it has been discussed whether APS derivation models as agreed during the Milan conference and briefly summarized above are sufficiently practical to convince stakeholders that they should be applied (17). Clearly, it is important that APS models should be capable of easy application and that, based on the chosen model, APS values should be established and put into practice. In doing this, we should acknowledge that not all desirable APS will be immediately achievable, but this approach will highlight which limitations of the current technology should be prioritized and solved. As described in Fig. 1, stakeholders are expected to work together in the field of APS to improve the contribution of laboratory medicine to patient care.

THE ROLE OF METROLOGY IN DETERMINING APS BASED ON REQUIRED CLINICAL CLASSIFICATION PERFORMANCE

When choosing which APS model should be employed, the most important aspect to consider is the intended use of the analyte. The most challenging context is when an analyte is used in different clinical situations, requiring tailor-made APS for each use of the measurand in a specific clinical decision.

One example of such a measurand is total bilirubin (TBIL). TBIL reflects the total amount of unconjugated, mono- and di-glucuronide conjugated and albumin-bound bilirubin in the blood. The intended clinical applications can be categorized into three areas: (a) diagnosis of liver disease; (b) diagnosis of hemolytic disease; and (c) diagnosis and monitoring of neonatal jaundice. While in the first two clinical conditions TBIL is used in conjunction with other important parameters (such as liver enzymes, alkaline phosphatase, and haptoglobin), for the latter clinical condition TBIL is the sole parameter on which treatment decisions are made. Depending on risk factors of the neonate, age-specific medical decision limits—the Buthani normogram—are globally used to guide initiation and monitoring of phototherapy and/or exchange transfusion (23).

Given the pivotal role of bilirubin in clinical decision making in neonatal jaundice, an APS based on this clinical outcome is clearly indicated (21). However, in practice, clinical outcome studies on the impact of neonatal bilirubin measurements are not available. Consequently, criteria derived from the BV model are most commonly applied, as this approach is preferred over the state-of-the-art model and is generally sufficient for clinical evaluation of hemoglobin breakdown and liver/bile duct function (reviewed in (24)). The EFLM database mentions desirable and optimal criteria for an expanded MU of 20.2% and 10.1%,

respectively, based on BV (25,26). However, BV-derived data expressed in the EFLM database and the underlying literature are based on adult populations and thus are not suitable for neonates with hyperbilirubinemia (27). CLIA has updated its performance specification in terms of total error allowable (TEa) for bilirubin to 20% in its 2025 External Quality Assessment (EQA) requirements without reference to the evidence source on which their requirements are based (28). Since, as previously shown (29,30), misclassification based on inaccurate TBIL results may have direct clinical consequences, an “expert opinion”-based TEa APS of 5% was set by the SKML ((Dutch) Foundation for external quality assessment in medical laboratory diagnostics) in the Netherlands, until an outcome-based APS is possible.

Recent advances in data processing and simulation, including the newly published “APS calculator” by Çubukçu et al. (31), offer opportunities to perform indirect clinical outcome studies, enabling the derivation of more appropriate APS for such cases. The APS simulator is a user-friendly web-based tool, which can be used to simulate the re-analysis of identical samples “in silico.” It assesses how MU affects the agreement between original and simulated results, using clinical decision limits as a reference. To ensure successful application of this tool, several factors must be taken into account. First, the input data set used is critical, as the data distribution directly influences the outcome. Specifically, if a greater number of data points cluster around a medical decision limit (MDL), there is an increased likelihood of values falling on either side of the MDL. Also, the data set used is assumed to reflect “the real concentration,” and therefore the results used need to be as unbiased as possible. Second, the MDL definition is very important. In neonatal hyperbilirubinemia, MDLs represent a continuum rather than a single fixed value. They vary depending on risk factors and differ for phototherapy and exchange

transfusion. This variability makes selecting an appropriate MDL for the APS tool challenging. Fortunately, the APS tool includes a feature that allows incorporation of multiple MDLs. However, when using multiple MDLs, the overall required APS tends to be much tighter compared to using a single MDL. This is because the likelihood of a result falling beyond at least one of the MDLs increases significantly when multiple MDLs are considered. Lastly, the tool generates minimal, desirable, and optimal APS, somewhat arbitrarily chosen for 90%, 95%, and 99% agreement (resulting in 10%, 5%, and 1% misclassification, respectively). However, before assuming, for example, 5% misclassification as “desirable,” relevant clinicians (in this case neonatologists) should be surveyed to establish the allowable misclassification definition.

The APS calculator was applied to neonatal bilirubin measurements using one year’s clinical data for neonates less than 7 days of age from a general hospital and an academic hospital. Both data sets were generated using assays with bias within $\pm 5\%$ of the reference method value. APS were calculated for both a single MDL and multiple MDLs. Using a single MDL of 12.9 mg/dL [220 $\mu\text{mol/L}$], the desirable APS was around 10% in an academic hospital setting and was more permissible compared to the APS obtained from data from a general hospital setting (APS of 4.5%), likely because data distribution in the latter was closer to the MDL. On the other hand, incorporating multiple MDLs resulted in unfeasibly stringent desirable APS values for both hospital settings—sometimes below 2% to keep misclassification at 5% level (data not published).

The key question of “What level of clinical misclassification is acceptable?” remains. For this, neonatologists should be surveyed. Since the consequences for neonates can vary depending on the clinical context—such as initiating or stopping phototherapy, performing an exchange transfusion, or whether neonatal care is provided in a hospital or at home—it is probably challenging to provide a general answer.

Other questions to be answered include how to practically implement the APSs once they have been established. These APS set the allowable targets for MU, which rely on bias that has been corrected or at least is within acceptable limits (3). However, substantial bias has been demonstrated in methods (24,29,30). The successor to the APS calculator, the APS simulator, is able to differentiate between allowable bias and allowable imprecision, effectively addressing the bias present in assays (32). However, the issue of the bias problem will be solved only by implementing metrological traceability of TBIL results to well-defined higher-order reference materials (24,27).

THE ROLE OF METROLOGY IN THE IDENTIFICATION OF EXCHANGEABILITY SUITABLE FOR EQUIVALENT INTERPRETATION

From the beginning of this century, we have shifted toward a new paradigm of automation, digitalization, and cross-enterprise data exchange. This interoperability of data between machines and software systems has had, and still has, its difficulties and will require additional arrangements that will ultimately lead to further standardization. As mentioned, interoperability has been highly implemented within healthcare organizations; however, in relation to data exchange between organizations, both nationally and internationally, there are still great hurdles to be overcome. The medical care system is changing worldwide, costs are ever increasing and care is becoming more complex. Therefore, governments and medical care providers are promoting and encouraging electronic health record (EHR) information exchange and decision support and artificial intelligence (AI) to relieve the increasing workload.

As medical information in the form of diagnostic results can already be exchanged from the analyzer to the electronic record, the basis of

interoperability, focus has been given to technical standardization. Syntax standards like Health Level Seven (HL7), American Society for Testing and Materials (ASTM) or Electronic Data Interchange for Administration, Commerce and Transport (EDIFACT) provide the backbone of this form of data exchange. However, data shipped via these technical standards is still not information if crucial metadata is missing for correct interpretation.

For that purpose, it is necessary to define the metadata set that is crucial for correct interpretation and thus correct use of diagnostic data in a clinical context. Information models for diagnostic data are currently defined and refined to meet the standards for correct information exchange of diagnostic results. However, to make these models universally understandable, moving from human-readable and standardized toward machine-readable and used throughout applications and algorithms, we need semantic standards. These semantic standards, like LOINC (33,34) (Logical Observations Identifier Names and Codes, governed by the Regenstrief Institute, United States), the NPU coding system (Nomenclature, Properties and Units) (35), UCUM (Unified Codes for Units of Measure) or SNOMED-CT (36) (Systematized Nomenclature of Medicine—Clinical Terms, SNOMED International), can be used to standardize diagnostic information, according to the diagnostic exchange information model. In this way, the LOINC-ID can be used to universally transcribe the analyte detected, UCUM to transcribe the unit of measurement, and SNOMED CT to define additional crucial metadata like specimen type and source. This information model is, in some countries, already used for cross-enterprise medical diagnostic information exchange. However, there is no general, widespread consensus as yet and laboratory specialists should be involved in the semantic standardization and configuration of the electronic systems used to ensure the correct mapping. Due to the fact that there are no data standardization protocols (i.e.,

laboratory semantic data exchange standards) and due to the lack of centralized governance on this important topic, medical diagnostic information exchange remains problematic. Furthermore, although LOINC has a slot available for the method used, none of the semantic standards adopt methodology, calibration, and/or metrological traceability in the code. However, governments now increasingly try to accelerate medical information exchange nationally and internationally.

In case of the EU, the EHDS initiative makes use of such a data model to enforce and enable medical diagnostic data exchange for patients for primary and secondary use. Therefore, we need to align our efforts to ensure that diagnostic data is exchangeable, comparable if possible, and interpretable. This is a duty for us as medical laboratory specialists that requires more focus on the post-analytical phase, because data with the correct context would become information. Additionally, machine-readable codes should be the basis of decision support, algorithms, and the specialized-use case for AI. If diagnostic data are not correctly standardized by means of machine-readable codes (semantic standardization), incorrect pre- or post-diagnostic data will devalue efforts made to improve decision support.

However, exchangeability and semantic standardization do not guarantee compatibility of results and correct interpretation by the receiving party. As an example, ferritin data with the LOINC ID 2276-4 can and eventually will be merged into a receiving EHR with internal ferritin data. However, based on the measuring system used, results may not be comparable (37). This is caused by the lack of standardization due to the availability to IVD manufacturers of different generations of WHO reference materials for ferritin with different metrological traceability options between these generations (2,38). In practice, this situation results in inconsistency and potential medical misinterpretation of ferritin results from different commercial IVD systems, despite identical LOINC codes.

For correct data processing, understanding and interpretation of the method of analysis by the receiving party can be crucial to allow appropriate presentation in electronic records. Additionally, for secondary use of diagnostic information, the analytical methods employed are a crucial and, in many cases, missing element, which can introduce research bias. Therefore, we need to include the method of analysis, supplier, and calibrator information into the information model for correct exchange of data from laboratory results (shown in Fig. 2). This will enrich our diagnostic data and guarantee the quality of internal and exchanged data. Furthermore, it will enable laboratory professionals to easily compare data across healthcare systems to evaluate methodological differences between measurement results and support a “big data analysis” approach for EQA and future harmonization efforts. However, to achieve these goals, coordination of coding and presentation of diagnostic information must be promoted by laboratory societies, laboratory professionals, semantic standard governance organizations, national governments, IVD manufacturers, and electronic health record (EHR) software developers.

METROLOGY AS A KNOWLEDGE DOMAIN WITH IMPORTANCE FOR ALL DISCIPLINES OF LABORATORY MEDICINE

Laboratory medicine comprises various medical disciplines in addition to clinical chemistry, including microbiology and pathology, among others. At first, it might seem that the majority of laboratory tests in these disciplines are qualitative tests, because of the dichotomic reporting format used for them. For instance, in the case of microbiology, is there evidence of infection (“yes or no”), can a microorganism be detected (“yes or no”), and, if so, which species is it and to which drugs is it susceptible. However, although often hidden,

quantification is often a crucial aspect in qualitative examinations as well. It is important not only for the determination of the limit of detection (LOD) or decision point, but also for determination of the variation of the measurement procedure, and thus for the reliability of the test. It could be argued that all quantitative examinations are in fact quantitative tests with a single threshold, deciding between 2 dichotomic results. In addition, quantification of microorganisms, or antibodies against them, is often also important for clinical decisions related to the treatment of the patient. Examples are the determination of the viral load in patients with HIV, the colony-forming units (CFU) for bacterial infections and the extent of parasitemia for *Plasmodium falciparum* infections. These quantitative examinations are, for instance, used to assess the severity of disease and/or the effect of treatment. Furthermore, examination of the growth rate of bacteria or fungi is critical in the antibiotic susceptibility testing that is used to determine the optimal antibiotic drug for treatment.

Metrological traceability to the intended measurand in medical microbiology faces the same challenges as those in clinical chemistry and has additional problems to address. In microbiology, the measurand is either the microorganism or specific antibodies against microorganisms. If variants exist of a certain microorganism that only differ from each other in very subtle aspects, these variants can still be seen as different measurands. In such cases it is essential to know which methods are able to detect which variants. In some cases, it will be important not to miss any variant, whereas in other cases it will be essential to distinguish between variants with different pathogenicity.

In contrast to most measurands in clinical chemistry, microorganisms and the specific antibodies against them cannot be defined chemically as they are complex and diverse. Furthermore, microorganisms are continuously changing (evolving by replication and selection by host immunity). Hence, the definition of the measurand cannot

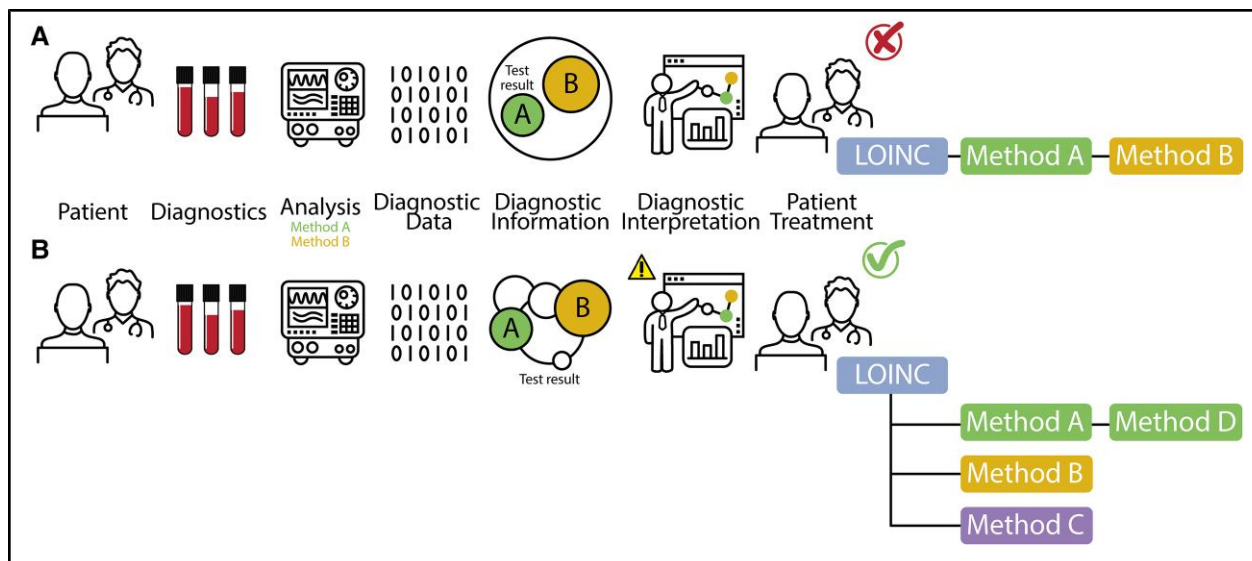


Fig. 2. (A), Same LOINC-ID, no metadata relating to methodology in exchanged information provided. A risk arises from presentation of aggregated results that could result in misinterpretation; (B), By providing discrete method grouping information (by using, for example, SNOMED-CT) via post-coordination software, systems capture methodology information for secondary research purposes, leading to correct data aggregation. The common term “method” reflects the stratifying factor (e.g., measurement procedure, instrument, or higher-order traceability anchor) that causes results (A and B) to be able to be grouped together (e.g., in A) or interpreted distinctly (e.g., in B).

be static and sometimes a purified measurand is not available (e.g., in a case where a microorganism cannot be cultured in vitro). Therefore, reference materials for microbiological tests are often lacking, which substantially hampers the implementation of metrology in microbiology. However, this is not a reason to disregard the documentation of metrological traceability. One even might argue that the lack of reference materials makes it even more important to document the exact status of the measurand, including which variants are recognized and which not.

The lack of reference materials in combination with differences in test principle (e.g., detection of a microorganism by plaque formation, microscopy, CFU, or nucleic acid amplification test (NAAT)) can result in a large variation in outcome of distinct test types (e.g., due to variation of the LOD), even between results of laboratories that use the same type of assay (e.g., due to differences

in the pre-analytic process). This variation between assays and laboratories is demonstrated by EQA schemes (39,40). In most cases these variations have limited clinical consequences as the results are often translated into a dichotomous result interpretation (“positive or negative”); in these circumstances, only measurand concentrations around the decision value are likely to have a variable result interpretation. In cases where not all assays have a detection limit allowing detection of relevant low concentrations of the measurand, EQA using commutable samples will help to identify such differences in LOD. Apart from sensitivity, selectivity can also be different between methods. If differences between methods result in relevant differences in the recognition of variants of the measurand, appropriate EQA will aim to identify such differences.

Correct interpretation of dichotomous results in the context of infectious diseases relies on

the awareness by requesting physicians of these possible sources of variation. Laboratory specialists have a responsibility to ensure awareness of relevant knowledge obtained with the help of EQA data among their requesting clinical partners.

In conclusion, metrological traceability and APS in medical disciplines other than clinical chemistry face the same challenges but to a greater extent as the measurand is often not accurately defined.

CONCLUSIONS

Metrological traceability and MU evaluation are key elements in the knowledge domain of laboratory specialists. This knowledge is vital to set

clinically relevant APS and the required specificity based on the intended population, acceptable clinical decision performance, and misclassification risks and consequences. A proper and unequivocal definition of the measurand and how this relates to the intended analyte is vital for understanding the basics of metrology and the correct implementation thereof. This will ensure that clinicians can make consistent decisions based on the same initial tests by acquiring comparable data presented as equivalent results. We urge (inter)national societies to become increasingly aware that their members need to master the skills involved by redesigning the education system in order to make laboratory professionals as futureproof and fit for purpose as the laboratory services for which they are responsible (3,41).

Nonstandard Abbreviations: IVD, in vitro diagnostics; CE, Conformité Européenne; LDT, laboratory-developed test; EHDS, European Health Data Space; EHR, electronic health record; EU, European Union; Hb, hemoglobin; MU, measurement uncertainty; LDL-c, low-density lipoprotein cholesterol; Lp(a), lipoprotein(a); NCEP, National Education Cholesterol Program; CVD, cardiovascular disease; PCSK9, proprotein convertase subtilisin/kexin type 9; EAS, European Atherosclerosis Society; EFLM, European Federation of Clinical Chemistry and Laboratory Medicine; apo B, apolipoprotein B; AACC, American Association for Clinical Chemistry; ISD, immunosuppressive drug monitoring; AT, antithrombin; APS, analytical performance specifications; BV, biological variation; TBIL, total bilirubin; TEa, total error allowable; EQA, external quality assessment; SKML, (Dutch) Foundation for external quality assessment in medical laboratory diagnostics; MDL, medical decision limit; EHR, electronic health record; AI, artificial intelligence; HL7, Health Level Seven; ASTM, American Society for Testing and Materials; EDIFACT, Electronic Data Interchange for Administration, Commerce and Transport; LOINC, Logical Observations Identifier Names and Codes; NPU, Nomenclature, Properties and Units; UCUM, Unified Codes for Units of Measure; SNOMED-CT, Systematized Nomenclature of Medicine—Clinical Terms; LOD, limit of detection; CFU, colony-forming units; NAAT, nucleic acid amplification test.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Marith van Schroyen Lantman (Conceptualization-Equal, Writing—original draft-Equal), Christa Cobbaert (Conceptualization-Supporting, Writing—original draft-Equal), Mauro Panteghini (Conceptualization-Supporting, Writing—original draft-Equal), Miranda van Berkel (Conceptualization-Supporting, Writing—original draft-Equal), Ruben Smeets (Conceptualization-Supporting, Writing—original draft-Equal), Jaap van Hellemond (Conceptualization-Supporting, Writing—original draft-Equal), and Marc Thelen (Conceptualization-Lead, Writing—original draft-Equal)

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: None declared.

Disclosures: M.H.M. Thelen is an associate editor for *The Journal of Applied Laboratory Medicine*, ADLM.

Role of Sponsor: No sponsor was declared.

REFERENCES

1. Thelen M. An essential contribution to the clinical chemistry syllabus. *J Appl Lab Med* 2021;6:341–3.
2. Panteghini M. An improved implementation of metrological traceability concepts is needed to benefit from standardization of laboratory results. *Clin Chem Lab Med* 2025;63:270–8.
3. Panteghini M, Krintus M. Establishing, evaluating and monitoring analytical quality in the traceability era. *Crit Rev Clin Lab Sci* 2025;62:148–81.
4. International Organization for Standardization (ISO). ISO17511:2020: In vitro diagnostic medical devices — Requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples. 2nd Ed. Geneva (Switzerland): ISO; 2020.
5. Miller WG, Greenberg N. Harmonization and standardization: where are we now? *J Appl Lab Med* 2021; 6:510–21.
6. Panteghini M, Camara JE, Delatour V, Van Uytvanghe K, Vesper HW, Zhang T. Feasibility of metrological traceability implementation using the joint committee on traceability in laboratory medicine database entries including the fulfillment of “fit-for-purpose” maximum allowable measurement uncertainty. *Clin Chem* 2024;70:1321–33.
7. van Schroyen LM. How measurement uncertainty impedes clinical decision-making [Dissertation]. Nijmegen (the Netherlands): Radboud University; 2024. 185 p.
8. Van der Burgt YFM, Cobbaert CM. Proteoform analysis to fulfill unmet clinical needs and reach global standardization of protein measurands in clinical chemistry proteomics. *Clin Lab Med* 2018;38:487–97.
9. Neumann JT, de Lemos JA, Apple FS, Leong DP. Cardiovascular biomarkers for risk stratification in primary prevention. *Eur Heart J* 2025;46:3823–43.
10. Cobbaert CM. Implementing cardiovascular precision diagnostics: laboratory specialists as catalysts? *Ann Clin Biochem* 2023;60:151–154.
11. Otvos JD. Our own inconvenient truth: LDL cholesterol is a flawed measure of LDL. Time to acknowledge the problem and its clinical implications. *FATS Newsletter LVD* 2009;XXIII:16–20.
12. Langlois MR, Chapman MJ, Cobbaert C, Mora S, Remaley AT, Ros E, et al. Quantifying atherogenic lipoproteins: current and future challenges in the era of personalized medicine and very low concentrations of LDL cholesterol. A consensus statement from EAS and EFLM. *Clin Chem* 2018;64:1006–33.
13. Contois JH, McConnell JP, Sethi AA, Csako G, Devaraj S, Hoefner DM, Warnick GR. Apolipoprotein B and cardiovascular disease risk: position statement from the AACC lipoproteins and vascular diseases division working group on best practices. *Clin Chem* 2009;55: 407–19.
14. Diederiks NM, van der Burgt YEM, Ruhaak LR, Cobbaert CM. Developing an SI-traceable Lp(a) reference measurement system: a pilgrimage to selective and accurate apo(a) quantification. *Crit Rev Clin Lab Sci* 2023; 60:483–501.
15. Kruijt, M. On the diversity of antithrombin proteoforms: the role of a diagnostic mass spectrometry-based test for antithrombin deficiency. <https://hdl.handle.net/1887/4175625> (Accessed January 2025).
16. Brunet M, van Gelder T, Åsberg A, Haufrøid V, Hesselink DA, Langman L, et al. Therapeutic drug monitoring of tacrolimus-personalized therapy: second consensus report. *Ther Drug Monit* 2019;41:261–307.
17. Panteghini M. What the Milan conference has taught us about analytical performance specification model definition and measurand allocation. *Clin Chem Lab Med* 2024;62:1455–61.
18. Ceriotti F, Fernandez-Calle P, Klee GG, Nordin G, Sandberg S, Streichert T, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM strategic conference. *Clin Chem Lab Med* 2017;55:189–94.
19. Smith AF, Shinkins B, Hall PS, Hulme CT, Messenger MP. Toward a framework for outcome-based analytical performance specifications: a methodology review of indirect methods for evaluating the impact of measurement uncertainty on clinical outcomes. *Clin Chem* 2019;65:1363–74.
20. Krintus M, Panteghini M. Judging the clinical suitability of analytical performance of cardiac troponin assays. *Clin Chem Lab Med* 2023;61:801–10.
21. Borrillo F, Panteghini M. State-of-the-art model for derivation of analytical performance specifications: how to define the highest level of analytical performance technically achievable. *Clin Chem Lab Med* 2024;62: 490–1496.
22. Cattaneo D, Panteghini M. Analytical performance specifications for measurement uncertainty in therapeutic monitoring of immunosuppressive drugs. *Clin Chem Lab Med* 2024;62:e81–3.
23. Maisels MJ, Bhutani VK, Bogen D, Newman TB, Stark AR, Watchko JF. Hyperbilirubinemia in the newborn infant > or =35 weeks' gestation: an update with clarifications. *Pediatrics* 2009;124:1193–8.
24. Hulzebos CV, Camara JE, van Berkel M, Delatour V, Lo SF, Mailloux A, et al. Bilirubin measurements in neonates: uniform neonatal treatment can only be achieved by improved standardization. *Clin Chem Lab Med* 2024;62: 1892–903.
25. EFLM. EFLM Biological Variation Database. <https://biologicalvariation.eu/> (Accessed September 2025).
26. Panteghini M, Braga F, Camara JE, Delatour V, Van Uytvanghe K, Vesper HW, Zhang T, et al. Optimizing available tools for achieving result standardization: value added by Joint Committee on Traceability in Laboratory Medicine (JCTLM). *Clin Chem* 2021;67: 1590–605.
27. Panteghini M, Miller WG, Wielgosz R. Time to refresh and integrate the JCTLM database entries for total bilirubin: the way forward. *Clin Chem Lab Med* 2025;63:e73–5.

28. US Department of Health and Human Services. CLIA proficiency testing regulations related to analytes and acceptable performance. <https://www.govinfo.gov/content/pkg/FR-2022-07-11/pdf/2022-14513.pdf> (Accessed September 2025).
29. Oostendorp M, Ten Hove CH, van Berkel M, Roovers L. A significant increase in the incidence of neonatal hyperbilirubinemia and phototherapy treatment due to a routine change in laboratory equipment. *Arch Pathol Lab Med* 2024;148:e40–7.
30. Thomas DH, Warner JV, Jones GRD, Chung JZY, Macey DJ, Screnci A, Ryan JB. Total bilirubin assay differences may cause inconsistent treatment decisions in neonatal hyperbilirubinaemia. *Clin Chem Lab Med* 2022;60:1736–44.
31. Çubukçu HC, Vanstapel F, Thelen M, van Schrojenstein Lantman M, Bernabeu-Andreu FA, Meško Brguljan P, et al. APS calculator: a data-driven tool for setting outcome-based analytical performance specifications for measurement uncertainty using specific clinical requirements and population data. *Clin Chem Lab Med* 2023;62:597–607.
32. Çubukçu HC. Computer simulation approaches to evaluate the interaction between analytical performance characteristics and clinical (mis)classification: a complementary tool for setting indirect outcome-based analytical performance specifications. *Clin Chem Lab Med* 2025;63:1292–300.
33. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996;42:81–90.
34. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003;49:624–33.
35. Joint Committee on Nomenclature, Properties and Units (C-SC-NPU) of the IFCC and IUPAC; Pontet F, Magdal Petersen U, Fuentes-Arderiu X, Nordin G, Bruunshuus I, et al. Clinical laboratory sciences data transmission: the NPU coding system. *Stud Health Technol Inform* 2009;150:265–9.
36. National Library of Medicine. Overview of SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html (Accessed September 2025).
37. Braga F, Pasqualetti S, Frusciante E, Borrillo F, Chibireva M, Panteghini M. Harmonization status of serum ferritin measurements and implications for use as marker of iron-related disorders. *Clin Chem* 2022;68:1202–10.
38. Swinkels DW, van Schrojenstein Lantman M, Matlung HL, Weykamp C, Thelen M. Equivalence in clinical assessment of iron status requires ferritin assay standardisation before harmonisation of ferritin reference intervals. *Lancet Haematol* 2024;11:e721.
39. Schuurs TA, Koelewijn R, Brienen EAT, Kortbeek T, Mank TG, Mulder B, et al. Harmonization of PCR-based detection of intestinal pathogens: experiences from the Dutch external quality assessment scheme on molecular diagnosis of protozoa in stool samples. *Clin Chem Lab Med* 2018;56:1722–7.
40. Schutte AHJ, Koelewijn R, Ajjampur SSR, Leveck B, McCarthy JS, Mejia R, et al. Detection of soil-transmitted helminths and *Schistosoma* spp. by nucleic acid amplification test: results of the first 5 years of the only international external quality assessment scheme. *PLoS Negl Trop Dis* 2024;18:e0012404.
41. Cobbaert C. Time for a holistic approach and standardization education in laboratory medicine. *Clin Chem Lab Med* 2017;55:311–3.