Opinion Paper

Andrea Rita Horvath*, Katy J.L. Bell, Ferruccio Ceriotti, Graham R.D. Jones, Tze Ping Loh, Sally Lord and Sverre Sandberg, on behalf of the Task Group Analytical Performance Specifications based on Outcomes of the European Federation of Clinical Chemistry and Laboratory Medicine

Outcome-based analytical performance specifications: current status and future challenges

https://doi.org/10.1515/cclm-2024-0125
Received January 25, 2024; accepted May 18, 2024;
published online June 6, 2024

Abstract: Analytical performance specifications (APS) based on outcomes refer to how 'good' the analytical performance of a test needs to be to do more good than harm to the patient. Analytical performance of a measurand affects its clinical performance. Without first setting clinical performance requirements, it is difficult to define how good analytically the test needs to be to meet medical needs. As testing is indirectly linked to health outcomes through clinical decisions on patient management, often simulation-based studies are used to assess the impact of analytical performance on the probability of clinical outcomes which is then translated to Model 1b APS according to the Milan consensus. This paper discusses the related key definitions, concepts and considerations that should assist in finding the most appropriate methods for deriving Model 1b APS. We review the advantages and limitations of published methods and discuss the criteria for transferability of Model 1b APS to different settings. We consider that the definition of the clinically acceptable misclassification rate is central to Model 1b APS. We provide some examples and guidance on a more systematic approach for first defining the clinical performance requirements for tests and we also highlight a few ideas to tackle the future challenges associated with providing outcome-based APS for laboratory testing.

Keywords: analytical performance; clinical performance; laboratory test; outcomes

***Corresponding author: Andrea Rita Horvath**, Department of Chemical Pathology, New South Wales Health Pathology, Prince of Wales Hospital, Level 4 Campus Centre, Randwick, NSW2031, Sydney, Australia; School of Public Health, University of Sydney, Sydney, Australia; and Faculty of Medicine, University of New South Wales, Sydney, Australia, E-mail: Andrea.Horvath@health.nsw.gov.au
Katy J.L. Bell, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia
Ferruccio Ceriotti, Clinical Laboratory, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
Graham R.D. Jones, Faculty of Medicine, University of New South Wales, Sydney, Australia; and Department of Chemical Pathology, SydPath, St Vincent's Hospital, Darlinghurst, NSW, Australia
Tze Ping Loh, Department of Laboratory Medicine, National University Hospital, Singapore, Singapore
Sally Lord, School of Medicine, University of Notre Dame, Darlinghurst, New South Wales, Australia; and NHMRC Clinical Trials Centre, The University of Sydney, Camperdown, New South Wales, Australia
Sverre Sandberg, Norwegian Organization for Quality Improvement of Laboratory Examinations (NOKLUS), Haraldsplass Deaconess Hospital, Bergen, Norway; Norwegian Porphyria Centre, Department of Medical Biochemistry and Pharmacology, Haukeland University Hospital, Bergen, Norway; and Institute of Public Health and Primary Health Care, University of Bergen, Bergen, Norway

Introduction

In 2014 in Milan, the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) was dedicated to a single topic on how analytical performance specifications (APS) should be defined to guide various stakeholders involved in the delivery of laboratory service. As a result of the meeting three non-hierarchical models were proposed to replace the Stockholm criteria released in 1999: Model 1 APS based on outcomes, Model 2 APS based on biological variation and Model 3 APS based on the state of the art [1]. The Milan consensus group has also recognised the importance of the pre- and postanalytical phases of testing and encouraged users to expand quality specifications to the total testing process [1].

In this paper we focus on Model 1 APS and the related key definitions, concepts and considerations that should assist in finding the most appropriate methods for deriving APS which contribute to improved laboratory service that delivers patient benefit. We also review published methods for Model 1

APS and discuss the current status and future challenges of defining outcome-based APS for laboratory tests.

Key definitions and concepts

According to the Milan consensus, "APS are criteria that specify (in numerical terms) the quality required for analytical performance in order to deliver laboratory test information that would satisfy clinical needs for improving health outcomes" [1].

To satisfy clinical needs and improve health outcomes through laboratory testing, laboratory professionals generally tend to believe that tests that perform better analytically may lead to better patient outcomes. However, this is only true, if the better performing tests 1/improve the clinical performance of the test: i.e., they increase the rate of appropriate diagnoses and decrease the rate of missed diagnoses; and 2/this improves the clinical effectiveness of management decisions: i.e., they support appropriate treatment selection [2–4]. In the context of the above definition of APS, the primary focus is clinical performance to satisfy clinical needs; other health care related organisational or economic outcomes such as test safety, accessibility, convenience, turn-around time and costs may be important mediators for health outcomes but are not primary considerations and can be assessed separately [4].

To further dissect the actual meaning of the Milan definition, APS refer to how 'good' the analytical performance of a test needs to be to do more good than harm to the patient. For answering this question, a few more key concepts need clarification. How can we determine what 'good' analytical performance is? At this point it is important to highlight the difference between analytical performance goals and analytical performance requirements. The former is more aspirational in terms of the quality we ideally would like to achieve if we had better and more advanced technology; whilst analytical performance requirement is more of a pragmatic term to describe the analytical quality we must/should and can realistically achieve with existing technology. In the literature these two types of 'good' are often used interchangeably under the umbrella term of APS and that can lead to confusion by end users.

The other key concept or question is to whom 'good' is good enough. Analytical performance specifications are used by many stakeholders, including the IVD industry, the regulator, the health care purchaser, accrediting bodies, the external quality assurance (EQA) provider, the medical laboratory, the clinical staff including clinical guideline groups, and ultimately and indirectly the patient. The purpose of the APS for each stakeholder might be somewhat different and currently they all use different methods or criteria for defining what 'good' analytical performance means to them.

The relationship between the analytical and clinical performance of tests is reciprocally interdependent. By clinical performance/clinical validity we mean the ability of a test to provide information for its intended use about a health condition or state of interest in the relevant population [5]. This includes diagnostic accuracy (ability to correctly identify whether or not a condition is present; e.g. sensitivity, specificity, negative and positive predictive values) and prognostic accuracy (ability to predict whether or not an event will occur in the future; e.g. risk classification, concordance or c-statistic). Analytical performance of a measurand affects its clinical performance. However, without first setting the clinical performance requirements, it is difficult to define how good analytically the test needs to be to achieve the desired clinical performance. The clinical performance requirements may vary for the same test depending on how and where it is used and how its decision limit is established; e.g. for a triage test to rule out a condition, the test needs to have a cut-off that offers the highest possible diagnostic sensitivity; for diagnosis the test needs to have a cut-off that offers high specificity; or for prognosis a cut-off value can be set at a clinically significant or critical risk limit at which more intensive treatment or change in managing the condition is required.

According to the cyclical test evaluation framework proposed by our group in an earlier publication [6] there is an interplay between how good analytically a test is and how it can screen, diagnose, prognose or monitor a condition in a certain population. In other words, how good its clinical performance is to achieve the best outcome for patients. Crucial to all this interplay is the clinical pathway, i.e. how a test is used in making decisions about the management of the patient, which then more directly influences the patient's health outcome. Thus, the link between testing and health outcomes is almost always indirect [7] and is dictated by the clinical pathway, and the purpose and role and the significance of the test in clinical decision making [6].

Due to this indirect link, high quality analytical performance potentially leading to higher clinical performance by itself does not lead to better clinical action or patient compliance or more effective treatment and improved health outcomes. For example, if a test with better analytical quality and better clinical performance is applied to the wrong patient or clinical scenario, or at the wrong time in the course of treatment, then there will not be higher therapeutic success or better health outcomes. On the other hand, analytically less well performing tests that play a small part in a complex clinical pathway may not necessarily lead

to adverse or unfavourable outcomes [4]. Sometimes tests with relatively inferior analytical or clinical performance, but that are accessible in a rural or remote or under-resourced setting may achieve better health outcomes than better performing tests only offered in a central laboratory far removed from patient care. Point of care tests are an example, where less stringent APS are used [8]. Timely (albeit poorer quality) results may avoid irreversible patient harm due to delayed treatment, or loss to follow-up. This complexity makes setting universal APS that meet every stakeholder's medical needs particularly hard and complicated, if not impossible, or even counterproductive [4]. This again highlights the importance of understanding how a test is used in practice and what the consequences of testing are to the patient.

How can analytical performance specifications be developed according to Milan Model 1?

Model 1a APS can be addressed by clinical studies, ideally diagnostic randomised controlled trials (RCTs), that directly look at the impact of differing analytical performance on clinical outcomes. These types of RCTs are as yet unavailable and aspirational, so most people turn to Model 1b indirect (simulation-based) approaches that assess the impact of analytical performance on the probability of clinical outcomes by investigating the impact on medical decisions for patient management (see details below; [9]) as intermediates to patient health outcomes [2].

Most Model 1b studies use simulation modelling of the impact of analytical imprecision and bias on patient classification, or how medical decisions on, for example drug dosing [10], or treatment selection or advice could be impacted. Smith et al. have systematically reviewed the literature for these indirect approaches by investigating how they assessed the impact of test measurement uncertainty (by their definition and in frequency order it included random, systematic and total analytical error) on downstream clinical performance, and operational, and economic outcomes [11]. Fifty four percent of published studies investigated tests that were used for monitoring, 42 % for diagnosis or screening, and 9 % for prognosis across 4 clinical topics of diabetes mellitus, cardiovascular disease, cancer, and metabolic or endocrine disorders. They found various indirect, mostly statistical approaches in the literature; e.g. for HbA1c that used distributional analysis [12] or regression analysis [13]; for glucose, an error model simulation [14] and decision analytic models [15]; for aminoglycoside antibiotics, error grid/contour plots [10]; and for calcium, cost curve analysis [16]. Another approach, most commonly used in health technology assessment, is the so-called linked evidence approach that uses systematically reviewed evidence on diagnostic accuracy of a medical test and investigates its impact on clinical decision making and on the effectiveness of consequent treatment options [2, 17, 18]. Several studies linked the impact of improved imprecision to downstream outcomes, via a Markov cost-effectiveness model estimating, for example, the impact of glycaemic rates on cardiac events, and subsequent mortality and quality-adjusted life-year [19].

The review of methods by Smith et al. [11] has also identified a common analytical framework underpinning the various methods. It consisted of 3 key steps: (a) assignment of "*true*" test values; (b) calculation of *measured* test values (incorporating uncertainty due to bias and imprecision); and (*c*) calculation of the *impact* of discrepancies between the true and measured value on specified outcomes. Most studies suffered from defining the 'true' values of a measurand from empirical data (that includes bias and imprecision) or simulated data (which are not from any real-life source). Assessing bias would be most ideal using commutable (quality assurance) materials against a target that was set by a fully traceable reference measurement procedure. However, assessing the real impact of deviation from the 'true' value that can lead to misclassification in real life should be done against the actual method, including its analytical and preanalytical conditions, that was used to establish the clinical decision limit around which modelling of misclassification is investigated. Therefore, appropriately conducted correlation studies on patient samples against a carefully chosen reference are more suited for such modelling as these give a more realistic assessment on the impact of bias and imprecision on classification rates around guideline driven decision limits. Another problem with data sources used in modelling studies is that due to method changes and lot-to-lot variations the modelled analytical errors also vary over time and some methods that were used when certain decision limits were established are no longer available with the advancement of technology. Therefore, local transferability of data coming from such modelling studies should be critically assessed.

The majority of these simulation models have no direct evidence on the true impact of the modelled misclassification on medical decisions and consequent patient health outcomes. The clinician always interprets the result in the context of other specific clinical and diagnostic information and weights the strength of the abnormality found. Borderline results are not considered as strongly as marked abnormalities and in the absence of borderline zones, or

comments that grade the abnormality, many of the simulation studies are naive imitations of the more complex clinical decision making. Given that clinicians often use laboratory tests as adjuncts in their decision making, simulation models most probably overestimate the real-life impact of tests on patient well-being. For realistic Model 1b APS, the assumptions coming out of modelling therefore need to be tested in new studies using risk assessment or clinical audit or other forms of outcome assessment that can provide clinical evidence about the true consequences to patients.

To illustrate the complexity of deriving Model 1 APS, we have looked at the example of HbA1c [20] as one of the measurands that was allocated to Model 1 of the Milan consensus [21]. HbA1c plays an important role in and has well-defined decision limits for screening and diagnosing diabetes as well as for monitoring the progression and response of patients to treatment. We have reviewed publications that investigated how good the HbA1c test needs to be to meet medical needs [20]. We extracted the APS from 18 papers that used various Milan models. Eight of these papers used a Model 1b study of which 5 used simulation modelling, 2 statistical derivation and one used an international clinician survey that investigated at what level of change in HbA1c result clinicians would alter their management decision, and translated that to an APS. Three of the 8 Model 1b studies provided no numerical APS and the rest provided APS for analytical coefficient of variation (CVa) that ranged from 2 to 9 per cent [20]. Most papers that used statistical approaches or simulation models came up with an analytical performance specification for CVa between 2 and 5 per cent or simply provided the magnitude of misclassification rates at various levels of imprecision and bias or assumed that bias was zero. The one international physician survey from 6 countries used a case scenario to assess at what change in the HbA1c result general practitioners would change their clinical management when diagnosing or monitoring diabetes mellitus. Using their responses as the clinically significant difference the authors calculated CVa using the reference change value (RCV) concept [22] and came up with much higher APS for CVa of up to 9 per cent at 95 % probability and even a higher figure of CVa of 25 per cent at 80 % probability than the earlier mentioned approaches [9]. In that study it was also interesting to see that clinicians tolerated far less analytical error when HbA1c results were increasing, i.e. the patient was deteriorating, than when the HbA1c was decreasing, and patient was improving [9]. Given these published figures, it is truly hard for any stakeholder to decide which analytical performance is good enough for successful management of diabetic patients. Nevertheless, these findings also indicate that in some scenarios clinicians may tolerate more analytical error than what comes out of statistical models. An important recent contribution to this type of analysis comes from assessing the contribution of non-glucose effects on HbA1c, e.g. such as average red cell life-span [23]. This approach recognises the additional complexity of unknown variation in the relationship between the measured result and the clinical classification. This raises the importance of understanding the pathophysiology of the measurands and the potential to overemphasise analytical performance when there are other unaccounted factors for variation in the results.

Clinically significant change approach

Clinically significant difference in the test result corresponds to a change in disease risk, prognosis, or response to treatment that would mandate a change in the patient's management. The most frequently cited APS for HbA1c are based on the clinically significant difference between consecutive HbA1c results. This approach uses the reference change value (RCV) calculation that describes the statistical likelihood of significant differences in serial test results from an individual [22]. The current recommendation, based on clinicians' opinion and studies that evaluated the effectiveness of new treatments in terms of the degree of HbA1c change, is that a 0.5 % (NGSP unit) or 5 mmol/mol (IFCC unit) change of HbA1c is clinically significant both at the 6.5 % or 48 mmol/mol diagnostic threshold and at the therapeutic decision threshold of 7 % or 53 mmol/mol [24]. This RCV target at 95 % probability can be accurately achieved if the intra-laboratory CVa of the HbA1c method is 2 % [24].

Another recently published study also used a statistical approach based on guideline driven critical difference between consecutive HbA1c results, combined with considerations given to intraindividual biological variation (CVi) and common preanalytical errors, to arrive at a 'clinically acceptable analytical performance specification'(CAAPS) [25]. They provided CAAPS for both diagnostic (4.3 % in IFCC units and 2.6 % in NGSP units) and monitoring purposes (2.4 % in IFCC units and 1.6 % in NGSP units) at the above clinically relevant decision limits of HbA1c. Out of the two sets of criteria they recommended using the stricter monitoring criteria as a generic CAAPS for HbA1c. To achieve this CAAPS the authors recommended repeated sampling for HbA1c measurements for the diagnosis of diabetes mellitus [25].

Similar approach was used by Kilpatrick et al. to establish pragmatic APS for beta-hydroxybutyrate for the diagnosis and monitoring of diabetic ketoacidosis [26]. They defined two clinical performance requirements for the beta-hydroxybutyrate test, and that the analytical

performance of the assay should be good enough to reliably distinguish 1/four predefined diagnostic categories based on beta-hydroxybutyrate concentrations, and 2/a beta-hydroxybutyrate RCV of 0.5 mmol/L in response to therapy. In their study, similarly to Rotgers et al. [25], they also defined a significantly higher allowable CVa of <21.5 % (at bias assumed to be zero) for diagnostic discrimination between the 4 diagnostic categories at >99 % certainty than for meeting the above specified RCV criterion in the monitoring scenario. According to this study, to reliably detect a 0.5 mmol/L fall in beta-hydroxybutyrate concentration from a high value of 3 mmol/L the CVa of the assay needs to be 5 , 7 and 9 % at 99 –95–90 % probability, respectively [26].

The strength of the approaches using clinically significant difference as a clinical performance specification, compared to some other statistical methods, is that it is pragmatic and more realistically reflects on what clinicians would do if they followed guidelines. This approach often combines Model 1 and 2 by taking into account other inherent variations such as CVi and common preanalytical variations that are rarely considered in most APS published so far. The weakness of this approach is that the critical difference that sets the overall error budget from which the CVa is derived after subtracting variation components such as CVi or preanalytical variation is often based on clinical consensus or opinion of experts that is already influenced by the state-of-the art analytical performance of the measurand. This is, however, the case for all model 1b approaches since experts usually decide what an acceptable rate of misclassification is or how much difference between results is clinically significant for them to take action, and such judgment is always influenced by current standard of practice and current experience with test performance. This also implies that measures such as CAAPS, in fact, combine information from all 3 Milan models.

APS based on review of the literature

Several papers have been published recently that provide desirable and minimum standard measurement uncertainty (MU) figures as a measure of APS for a large number of common laboratory tests [27, 28]. For HbA1c the authors recommend using a desirable MU of 3 % that would lead to 2 % misclassification of diabetes and a minimum MU of 3.7 % that would lead to 3.7 % misclassification based on a paper by Nielsen et al. [29]. The limitation of such recommendation is that selection of one particular study for the APS may lead to selection bias as various modelling approaches have already come up with variable misclassification rates for the same measurand, and it is also unclear what misclassification rate is acceptable to various users of the APS.

Limitations of indirect approaches for Model 1b APS

For the moment we do not yet have accepted methodology for deriving Model 1b APS and published studies provide heterogeneous data that are difficult to synthesize. To assess the quality of such studies, we do not have a critical appraisal and meta-analysis tool that could be used to generate estimates for APS from papers using indirect approaches. The APS provided in most papers assume that long-term bias of a method is incorporated in the MU estimate and that any bias in test results has been controlled and preferably eliminated. This assumption does not realistically reflect routine laboratory services and already known biases between methods.

An important limitation of selecting APS from indirect studies is that they model misclassification in a certain population that may not be transferable to another population or setting (see further discussion of the impact a disease prevalence and spectrum later). Furthermore, the true impact of misclassification rates presented in various modelling studies has not been assessed in practice or in a clinical setting that differs from the study population and whether the APS derived from a certain study are clinically suitable or required. Therefore, most published Model 1b APS are based on untested assumptions, and some could be too stringent, others potentially too loose.

Hyohdoh et al. conducted a study that aimed to more realistically assess when physicians action a test result in real life using a statistical approach to analyse 65 million laboratory test data from 99,000 patients and ordering patterns in medical records [30]. Using repeat test intervals as indications for considering a test result abnormal enough to warrant further monitoring or actions and linking laboratory results to therapeutic decisions of initiating the prescription of specific medications, they derived so-called 'real-world clinical decision intervals' [30]. These 'real-life' decision intervals could also be translated to clinically acceptable deviations of laboratory results from reference intervals or decision limits – in other words values that could be used similarly to the clinically significant difference mentioned above, but that triggered action in real life, rather than being based on a few experts' opinion. This approach is linked to the concept of the clinical importance of misclassification. For example, wrongly classifying a patient with an HbA1c of 6.6 % as not having diabetes is not as important as to do so for a patient with a value of 7.2 %, where treatment

benefit is more likely. Clinical outcome studies also have uncertainties about the decision points, and "staged" misclassification rates (e.g. likely minor, moderate, major clinical effects) may be more relevant than a total number or proportion of patients misclassified.

Irrespective of the approaches used so far, the current status of Model 1b APS is that most assess the theoretical impact of imprecision and bias (or sometimes bias is assumed to be zero) on clinical classification. Often the published models do not directly translate to APS; they simply inform about potential classification errors at a certain degree of imprecision and bias in a certain modelled, often hypothetical population which may not always be comparable to the local population in which the APS is supposed to be used. Therefore, two key questions remain: 1/How much analytical error is tolerable without severely affecting disease classification, management decisions and health outcomes? 2/How transferable these estimates are to various patient populations and settings with prevalence of disease or disease spectrum that differs from the studies which modelled the impact?

Defining minimum acceptable clinical performance

What can be considered a clinically acceptable misclassification rate is the central component of Model 1b APS. To answer this question, we need clinical consensus and, even more importantly, we need to understand how good a test needs to be in terms of clinical performance before we can decide how good it needs to be analytically to meet those clinical performance expectations.

Target product profiles (TPPs) for medical tests would be one good example of how clinical performance requirements could be set *a priori* which is more often done for new biomarkers. Clinical performance specifications are a set of criteria that quantify the clinical performance a new test must attain to allow better health outcomes than current practice [5]. These describe the necessary properties of a new test to address an unmet clinical need. They often start with considering the end, i.e. the health outcome of testing. They are usually produced by regulators or health care purchasers or public health institutions to guide manufacturers in the development of 'fit for purpose' tests [31].

Most TPPs for diagnostic tests were developed for infectious diseases. For example, in the UK the Medical and Healthcare products Regulatory Agency (MHRA) set TPPs for point of care tests for the detection of acute SARS-CoV-2 infection in people with or without symptoms. The published TPP was based on the consensus of what is 'minimally acceptable' in the opinion of the IVD industry, healthcare professionals and medical device regulators in the UK. They have set desirable and acceptable clinical performance criteria for diagnostic sensitivity at ≥95 and ≥80 % for ruling in the infection, respectively; and for diagnostic specificity at ≥99 and ≥95 % for ruling out the infection, respectively. For more details the reader is referred to Target Product Profile: Point of Care SARS-CoV-2 detection tests – GOV.UK (www.gov.uk). One may argue that these TPPs also suffer from subjectivity of expert opinion and lack of agreed and transparent methodology for setting clinical performance specifications. Often, the clinical performance specification is set arbitrarily by statistical 'convention' around 80 , 90, 95 and 99 % statistical probability or confidence.

The paper by Lord et al. provides further guidance on a more systematic approach for defining test performance that meets clinical need [5]. Using decision analytic principles, this paper provides a 5-step practical guide for developing minimum clinical performance requirements that consider the trade-off between the benefits and harms of a test and use net benefit as a measure [32]. Unquestionably, the trade-off between benefit and harm is a value judgement and therefore it also requires expert consensus. This judgment may also vary between health-care settings due to economic and organisational considerations. As pointed out earlier, for meaningful estimates of clinical performance, diagnostic accuracy of a test and the consequences of true positive and false positive or true negative and false negative tests need to be assessed in a population and in a well-defined clinical pathway that closely reflects how the test is intended to be used in practice [5]. One classic and widely known example for such value judgment has been described by Than et al. [33]. Emergency medicine physicians from the US, Canada, Australia and New Zealand were surveyed at various professional conferences and internally for the acceptable risk of false negative Troponin results when patients present with symptoms of chest pain in the emergency department. The consequence of a false negative Troponin result was defined as a missed major adverse cardiac event (MACE) within 30 days after discharge. One thousand twenty-nine clinicians responded to the survey and 40 % accepted 1 % or higher false negative result. Fifty five percent accepted 0.5 % or higher miss rate. The investigators used these data to conclude "clinicians may expect diagnostic strategies for the assessment of suspected ACS to achieve a sensitivity of 99 % or higher for AMI or other MACE."[33].

This leads to the next key question: How transferable these estimates are to different patient populations and settings with prevalence of disease or disease spectrum that

differs from the studies which modelled the impact of analytical performance on classification rates? This is an often-overlooked aspect of modelling studies, even though we have clear evidence from multiple studies that disease misclassification rates depend on the prevalence/pretest probability of the condition. In addition to the impact of prevalence on the positive and negative predictive value of a test, sensitivity and specificity may also vary. For example, with increasing prevalence, lower specificity and higher sensitivity have been observed due to differences in disease spectrum in studies [34–36]. From the above it follows that APS apply to the actual population and the diagnostic pathway in which the clinical performance requirements were derived, and the comparability of the local and published population and clinical pathway need to be assessed before using the APS.

Analytical performance specification is a moving target

The various aspects mentioned above should be carefully considered before adopting any APS based on Model 1b studies. As already emphasised, APS should respond to clinical needs, but the clinical performance requirements are also highly dependent on patient population, disease spectrum and prevalence and existing practice in addition to many other pragmatic organisational, geographical, economic and sometimes societal considerations that often influence medical laboratory practice. Therefore, APS cannot always be fully harmonised or rigidly applied without taking into account the priorities and preferences of the local health care system.

Analytical performance specifications also have constantly moving targets not only because of rapid technological developments in the laboratory field but also due to changes in the ways conditions are managed and treated with newer medications which alter the needs and expectations of clinicians and patients. One good example to illustrate this point is the APS for LDL-cholesterol that is still the recommended primary target of lipid lowering therapy [37]. According to the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS), the treatment goals for patients at very high risk and high risk of atherosclerotic cardiovascular disease are <1.4 mmol/L (<55 mg/dL) and <1.8 mmol/L (<70 mg/dL), respectively [38]. The last APS of CVa <4 %, bias ≤4 % and total error ≤12 % were developed by expert consensus of the National Cholesterol Education Program (NCEP) for higher LDL-cholesterol decision limits that guided on-treatment goals and initiation of treatment at 2.6 mmol/L (100 mg/dL) and 4.9 mmol/L (190 mg/dL), respectively [39]. Cole et al. have recently demonstrated by modelling that using the NCEP APS for risk stratification based on calculated LDL-C, up to 10 % of cases would be misclassified into a different risk group, potentially leading to mismanagement. These risk groups ranged from LDL-cholesterol of <1.8 mmol/L (<70 mg/dL) to >4.9 mmol/L (>190 mg/dL). To reduce the rate of inappropriate risk stratification, authors proposed the adoption of tighter APS by reducing the NCEP allowable bias to ≤3 % and the imprecision to CVa <3 %, which, according to the state-of-the-art, are currently achievable by most lipid methods [40, 41]. They modelled the misclassification rate after lowering the APS and found a modest reduction by up to 10 % and, unsurprisingly, that reduction in proportional bias had a larger impact than reduction in imprecision on the number of cases misclassified. Again, stricter APS seem to be justified if the true misclassification rate is high, but it would be good to have firmer evidence from clinical practice whether improved analytical performance would truly lead to better patient management and outcomes.

Future challenges

Given all the complexities and shortcomings of existing methods we need to recognise that we still have significant gaps in our knowledge and lack the tools for better defining Model 1b APS that are universally implementable. Quoting Wytze Oosterhuis [42]: “The practice of the clinical laboratory is such, that it is impossible to describe performance specifications in a mathematically perfect model, and all models will be based on assumptions and can only approach complex reality. The challenge is to reach consensus on a model that is both useful and as less flawed as possible.”

How should we fill the gap until we have better studies, least flawed methodologies and consensus that tell us how good we need to be? The below is mostly personal opinion and we raise these points to generate more discussion and broader international collaboration of key stakeholders, and facilitate further research in this area. As Model 1b APS should respond to clinical needs, one of the first steps is that we as laboratory professionals start translating our analytical performance to clinical performance metrics and apply risk-based approaches [43] in order to quantify the impact of potential misclassification in our local population. Such an activity should include the assessment of the combined impact of analytical, biological and preanalytical variations, by defining and reporting a grey zone around discrete decision thresholds, and facilitate repeat testing, when necessary, before a diagnostic or treatment decision is made.

Currently we already have a few published and freely available but as yet unvalidated statistical and interactive

tools [44–46] and a large amount of real-life laboratory data that could help us inform our clinicians, guideline developers and the IVD industry about the impact of analytical performance on test accuracy and clinical decisions. These first steps would also offer an opportunity for more patient focused conversations between laboratorians, clinicians and the IVD industry and for a more concerted and globally better coordinated effort to improve the safety and quality of laboratory service in a way that better meets patients' medical needs.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interests.
Research funding: None declared.
Data availability: Not applicable.

References

1. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: consensus statement from the 1st strategic conference of the European federation of clinical chemistry and laboratory medicine. Clin Chem Lab Med 2015; 53:833–5.
2. Staub LP, Lord SJ, Simes RJ, Dyer S, Houssami N, Chen RYM, et al. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. BMC Med Res Methodol 2012;12:12.
3. di Ruffano FL, Hyde C, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. Br Med J 2012;344:e686.
4. Horvath AR, Bossuyt PMM, Sandberg S, StJohn A, Monaghan PJ, Verhagen-Kamerbeek WDJ, et al. For the 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; 53:841–8.
5. Lord SJ, StJohn A, Bossuyt PMM, Sandberg S, Monaghan PJ, O'Kane M, et al. For the test evaluation working group of the European federation of clinical chemistry and laboratory medicine. Setting clinical performance specifications to develop and evaluate biomarkers for clinical use. Ann Clin Biochem 2019;56:527–35.
6. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. For the test evaluation working group of the European federation of clinical chemistry and laboratory medicine. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 2014;427:49–57.
7. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol 2014;67:612–21.
8. Stavelin A, Sandberg S. Analytical performance specifications and quality assurance of point-of-care testing in primary healthcare. Crit Rev Clin Lab Sci 2023;61:164–77.
9. Skeie S, Nordin G, Oosterhuis WP, Araczki A, Horvath AR, Perich C, et al. Post-analytical external quality assurance of blood glucose and HbA1c: an international survey. Clin Chem 2005;51:1145–53.
10. Nguyen TA, Kirubakaran R, Schultz HB, Wong S, Reuter SE, McMullan B, et al. Analytical and non-analytical variation may lead to inappropriate antimicrobial dosing in neonates: an in silico study. Clin Chem 2023;69: 637–48.
11. 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.
12. Chai JH, Ma S, Heng D, Yoong J, Lim WY, Toh SA, et al. Impact of analytical and biological variations on classification of diabetes using fasting plasma glucose, oral glucose tolerance test and HbA1c. Sci Rep 2017;7:7.
13. Asberg A, Odsater IH, Carlsen SM, Mikkelsen G. Using the likelihood ratio to evaluate allowable total error—an example with glycated hemoglobin (HbA1c). Clin Chem Lab Med 2015;53:1459–64.
14. Boyd JC, Bruns DE. Quality specifications for glucose meters: assessment by simulation modeling of errors in insulin dose. Clin Chem 2001;47:209–14.
15. Boyd JC, Bruns DE. Effects of measurement frequency on analytical quality required for glucose measurements in intensive care units: assessments by simulation models. Clin Chem 2014;60:644–50.
16. Gallaher MP, Mobley LR, Klee GG, Schryver P. The impact of calibration error in medical decision making. Washington, DC: National Institute of Standards and Technology; 2004.
17. Trikalinos TA, Siebert U, Lau J. Decision-analytic modelling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009;29:E22–9.
18. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care 2013;29:343–50.
19. Breton MD, Hinzmann R, Campos-Nanez E, Riddle S, Schoemaker M, Schmelzeisen-Redeker G. Analysis of the accuracy and performance of a continuous glucose monitoring sensor prototype: an in-silico study using the UVA/PADOVA type 1 diabetes simulator. J Diabetes Sci Technol 2017;11:545–52.
20. Loh TP, Smith AF, Bell KJL, Lord SJ, Ceriotti F, Jones G, et al. Setting analytical performance specifications using HbA1c as a model measurand. Clin Chim Acta 2021;523:407–14.
21. 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.
22. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. Crit Rev Clin Lab Sci 1989;27:409–37.
23. Weykamp C, Siebelder C, Lenters E, Slingerland R, English E. The risk of clinical misinterpretation of HbA1c: modelling the impact of biological variation and analytical performance on HbA1c used for diagnosis and monitoring of diabetes. Clin Chim Acta 2023;548:117495.
24. Little RR, Rohlfing CL, Sacks DB for the National Glycohemoglobin Standardization Program (NGSP) Steering Committee. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. Clin Chem 2011;57:205–14.
25. Rotgers E, Linko S, Theodorsson E, Kouri TT. Clinical decision limits as criteria for setting analytical performance specifications for laboratory tests. Clin Chim Acta 2023;540:117233.

26. Kilpatrick ES, Butler AE, Atkin SL, Sacks DB. Establishing pragmatic analytical performance specifications for blood beta-hydroxybutyrate testing. Clin Chem 2023;69:519–24.
27. Braga F, Panteghini M. Performance specifications for measurement uncertainty of common biochemical measurands according to Milan models. Clin Chem Lab Med 2021;59:1362–8.
28. Braga F, Pasqualetti S, Borrillo F, Capoferri A, Chibireva M, Rovegno L, et al. Definition and application of performance specifications for measurement uncertainty of 23 common laboratory tests: linking theory to daily practice. Clin Chem Lab Med 2023;61:213–23.
29. Nielsen AA, Petersen PH, Green A, Christensen C, Christensen H, Brandslun I. Changing from glucose to HbA1c for diabetes diagnosis: predictive values of one test and importance of analytical bias and imprecision. Clin Chem Lab Med 2014;52:1069–77.
30. Hyohdoh Y, Hatakeyama Y, Okuhara Y. A simple method to identify real-world clinical decision intervals of laboratory tests from clinical data. Inform Med Unlocked 2021;23:100512.
31. Cocco P, Ayaz-Shah A, Messenger MP, West RM, Shinkins B. Target Product Profiles for medical tests: a systematic review of current methods. BMC Med 2020;18:119.
32. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.
33. Than M, Herbert M, Flaws D, Cullen L, Hess E, Hollander JE, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the emergency department? A clinical survey. Int J Cardiol 2013;166:752–4.
34. Leeflang MM, Rutjes AW, Reitsma JB, Hooft BPMM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185: E537–44.
35. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ 2016;353:i3139.
36. Murad MH, Lin L, Chu H, Hasan B, Alsibai RA, Abbas AS, et al. The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy. CMAJ 2023;195:E925–31.
37. Langlois MR, Nordestgaard BG, Langsted A, Chapman MJ, Aakre KM, Baum H, et al. Quantifying atherogenic lipoproteins for lipid-lowering strategies: consensus-based recommendations from EAS and EFLM. Clin Chem Lab Med 2020;58:496–517.
38. Mach F, Baigent C, Catapano AL, Koskinas KL, Casula M, Badimon L, et al. The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J 2020;41:111–88.
39. National Cholesterol Education Program (NCEP). Second report of the expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel II). Circulation 1994;89: 1333–445.
40. Cole J, Sampson M, van Deventer HE, Remaley AT. Reducing lipid panel error allowances to improve the accuracy of cardiovascular risk stratification. Clin Chem 2023;69:1145–54.
41. Cobbaert CM. Editorial: implementing cardiovascular precision diagnostics: laboratory specialists as catalysts? Ann Clin Biochem 2023; 60:151–4.
42. Oosterhuis WP. Analytical performance specifications in clinical chemistry: the holy grail? J Lab Precis Med 2017;2:78.
43. Schmidt RL, Straseski JA, Raphael KL, Adams AH, Lehman CM. A risk assessment of the jaffe vs enzymatic method for creatinine measurement in an outpatient population. PLoS One 2015;10: e0143205.
44. Chatzimichail T, Hatjimihail AT. A software tool for exploring the relation between diagnostic accuracy and measurement uncertainty. Diagnostics 2020;10:610.
45. Loh TP, Markus C, Lim CY. Impact of analytical imprecision and bias on patient classification. Am J Clin Pathol 2024;161:4–8.
46. Çubukçu HC, Vanstapel F, Thelen M, van Schrojenstein Lantman M, Bernabeu-Andreu FA, Brguljan PM, et al, On behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group Accreditation, ISO/CEN Standards (WG-A/ISO). 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.