IFCC Paper

Andrea R. Horvath, Sverre Sandberg, Marc H.M. Thelen, Mauro Panteghini, Hikmet Can Çubukçu, Emmi Rotgers, Eric Kilpatrick, Katy J.L. Bell, Graham R.D. Jones and Elvar Theodorsson*

Model 1b analytical performance specifications – report on the Prague 2025 workshop

https://doi.org/10.1515/cclm-2026-1024
Received June 25, 2026; accepted July 27, 2026;
published online August 18, 2026

Abstract: The main purpose of establishing APS is to ensure that test performance aligns with their specific clinical goals. Analytical performance specifications serve multiple functions: as acceptance criteria to verify that *in vitro* diagnostic medical devices (IVD MD) meet regulatory and quality standards before market release; for external quality assurance assessments; for quality control, monitoring reagent lot-to-lot variability, and allowing preanalytical variation assessments in labs; and for guiding purchasing decisions by healthcare providers, reimbursement agencies, and both public and private insurers. Over the years, various systems and projects have been established to evaluate the current state of the art (Model 3) of IVD MD through external quality assurance (EQA). Different national and international EQA providers often use distinct acceptance criteria tailored to their organisational goals and local healthcare settings. Model 2, which relies on biological variation data, has been effectively supported by the European Federation of Clinical Chemistry and Laboratory Medicine's database and its interactive tools for calculating APS (www.biologicalvariation.eu). Nonetheless, debates continue about optimal APS setting methods and the most suitable models for different measurands and scenarios. To explore these issues further, workshops were held in Prague in 2023 and 2025. This report presents the expert presentations from the 2025 workshop, which exclusively focused on APS based on Model 1b.

Keywords: analytical performance specifications; laboratory tests; outcomes

Preface

Laboratory professionals have debated for decades how to define analytical performance specifications (APS) for various laboratory medicine applications. The main purpose of establishing APS is to ensure that test performance aligns with their specific clinical goals. Analytical performance specifications serve multiple functions: as acceptance criteria to verify that *in vitro* diagnostic medical devices (IVD MD) meet regulatory and quality standards before market release; for

***Corresponding author: Elvar Theodorsson**, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden, E-mail: elvar.theodorsson@liu.se
Andrea R. Horvath, Department of Chemical Pathology, New South Wales Health Pathology, Prince of Wales Hospital, Level 4 Campus Centre, Randwick, Sydney, NSW, Australia; School of Public Health, University of Sydney, Sydney, Australia; and Faculty of Medicine, University of New South Wales, Sydney, 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
Marc H.M. Thelen, Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen, The Netherlands; and Foundation for Quality Assurance in Medical Laboratory Diagnostics (SKML), Nijmegen, The Netherlands. https://orcid.org/0000-0003-1771-669X
Mauro Panteghini, Department of Laboratory Medicine, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Bydgoszcz, Torun, Poland. https://orcid.org/0000-0002-7147-3433
Hikmet Can Çubukçu, Department of Medical Biochemistry, Ministry of Health Sincan Training and Research Hospital, Ankara, Türkiye. https://orcid.org/0000-0001-5321-9354
Emmi Rotgers, Department of Clinical Chemistry, University of Helsinki, and HUSLAB, HUS Diagnostic Center, Helsinki University Hospital, Helsinki, Finland
Eric Kilpatrick, Manchester University NHS Foundation Trust, Manchester, UK
Katy J.L. Bell, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia. https://orcid.org/0000-0002-0137-3218
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

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

external quality assurance assessments; for quality control, monitoring reagent lot-to-lot variability, and allowing pre-analytical variation assessments in labs; and for guiding purchasing decisions by healthcare providers, reimbursement agencies, and both public and private insurers.

The fundamental principles and framework were established in 1999 at the Stockholm expert consensus meeting [1]. A review was conducted 15 years later in 2014 in Milan, resulting in the current three models: Model 1, which includes Model 1a based on direct outcomes and Model 1b based on indirect outcomes like how analytical performance influences medical decisions; Model 2, which is based on biological variation; and Model 3, which reflects the latest status of analytical techniques [2].

Over the years, various systems and projects have been established to evaluate the current state of the art (Model 3) of IVD MD through external quality assurance (EQA). Different national and international EQA providers often use distinct acceptance criteria tailored to their organisational goals and local healthcare settings. Model 2, which relies on biological variation data, has been effectively supported by the European Federation of Clinical Chemistry and Laboratory Medicine's database and its interactive tools for calculating APS (www.biologicalvariation.eu). Nonetheless, debates continue about optimal APS setting methods and the most suitable models for different measurands and scenarios. To explore these issues further, workshops were held in Prague in 2023 and 2025. This report presents the expert presentations from the 2025 workshop, which exclusively focused on APS based on Model 1b.

Introduction to the workshop (Andrea Rita Horvath)

The Eurachem Guide for the "Fitness for Purpose of Analytical Methods" states that "For an analytical result to be fit for its intended purpose, it must be sufficiently reliable so that any decision based on it can be taken with confidence" [3]. This workshop focuses on how to develop analytical performance specifications (APS) to ensure that a laboratory test or method is fit for its intended clinical use and supports safe and effective clinical decisions that benefit patients. This has been defined by the Milan consensus group as a Model 1b indirect approach for setting analytical performance specifications (APS) [4].

Laboratory test performance, covering analytical, pre-analytical, and postanalytical processes, and laboratory test results as a product of such a testing process, do not directly impact health outcomes. Clinical decisions based on test results along the overall patient management pathway do so by informing medical choices, which ultimately influence patients' health and wellbeing [4]. Multiple factors, beyond test results and their variations, affect individual patients' clinical pathways and the subsequent medical actions. These include differences in medical practice, the spectrum and severity of the disease, family history, comorbidities, the actual role and significance of the laboratory test in making management decisions, the potential need for additional clinical or imaging investigations, the choice of and adherence to and response to treatments, patient preferences, and associated costs. Many of these factors should be taken into account when assessing the overall and indirect impact of the test and its performance on patient outcomes [2]. This highlights the importance of understanding how the laboratory test is used in the clinical pathway, and the complexity of this approach makes studies of Model 1b APS quite intricate [2, 5].

Most studies published so far to define Model 1b APS have focused on the impact of tests on the classification of conditions, which, due to the above-mentioned indirect linkage, cannot be considered true health outcomes. Most studies used various error simulation approaches, which are summarised together with a framework for indirect outcome-based APS, by Smith et al. [6]. Any form of APS developed for a given test using any Milan model should demonstrate both internal and external validity. The internal validity of APS reflects the quality of the dataset and the reliability of the methods used for its development, and that the assumptions underlying the model reflect the usual or most typical clinical use of the test. External validity is also crucial, as it determines how well APS from a single study in a particular laboratory or clinical setting can be generalised to other contexts, including different patient populations with varying disease spectra and prevalence rates. While achieving high analytical performance is generally desirable, it does not automatically guarantee effective clinical decisions or patient adherence to treatment. Therefore, methods for evaluating the indirect effects of test results on patient outcomes should be pragmatic, flexible, representative of the intended use, and thoughtfully designed to be applicable in practice. Model 1b APS should be beneficial for the public, patients, laboratories, and clinicians, and act as a resource for manufacturers and regulators. This raises the question of whether Model 1b APS can be a universally adopted figure for all?

Similar to previous meetings on this topic held in Milan (2014) [2] and Prague (2023) [7], this workshop, returning to Prague in 2025, is expected to lead to conclusions on the most practical approaches for defining Model 1b APS. The objective of this workshop is also to summarise the key considerations, supplemented by a checklist of essential items, that would inform the design and conduct of future simulations

or other models investigating the indirect impact of test results on patient outcomes and translating those findings into clinically relevant and internally and externally valid APS. We also expect that this workshop provides guidance on the specific steps to follow when developing Model 1b APS by indirect methods and on its use by medical laboratories, clinical researchers, guideline developers, quality assurance organisations, manufacturers, and regulators.

Clinician surveys to define Model 1b APS (Sverre Sandberg)

Clinicians often assume that laboratory results have no measurement uncertainty. Therefore, evaluating clinicians' opinions on how variations in measurement results affect patient outcomes must be done through indirect questions about how they interpret and action test results in clinical practice. Responses to questions like "what difference between two results they will react to" can clarify what measurement uncertainty and, thus, APS clinicians consider in their practice [8–10]. This enables the investigation of the analytical quality clinicians assume they have, and the use of this to set performance requirements that reflect how clinicians and patients utilise the tests. The primary objective of this approach is to determine which APS is required to meet the expectations and needs of physicians across various medical specialities.

An example of such a question is the following: A 37-year-old man has suffered from a duodenal ulcer twice previously. Last month, his indigestion recurred, and it has become a great deal worse this last week. Five weeks ago, his blood haemoglobin concentration was 111 g/L. State how low his haemoglobin must be (at least) before you take action……. g/L.

The *strengths* of this method for establishing model 1b APS in this way are: 1/it is convenient to create questionnaires, collect results, and calculate the APS; 2/the resulting APS does not rely on the prevalence of the actual condition in the examined population since this is indirectly given in the questionnaire; and 3/the APS reflects how clinicians use the tests in practice. The *limitations* of this approach to establishing APS are that: 1/there is significant variation in opinions among different clinicians; 2/clinicians' views may vary depending on local and personal experience due to geography, intended population, and prevalence; 3/the APS is calculated using the within-subject biological variation and is influenced by the state of art quality; and 4/the APS based on clinicians' opinions can only be used for measurands that are essential in diagnosing and monitoring specific conditions.

The impact of analytical performance on clinical decision making: from as good as it gets to as good as needed (Marc Thelen)

Clinical laboratories have traditionally focused on performance criteria related to measurements. However, the importance of measurement results in achieving optimal patient outcomes is increasingly gaining attention [11–14].

The main factors affecting the relationship between analytical performance and clinical diagnostic performance are: 1/the analytical performance, measured as bias and imprecision; 2/the distribution of patient results around the decision limits; and 3/how well the decision limits fit the intended clinical application.

To determine whether analytical performance meets the diagnostic requirements for a specific clinical application, an active dialogue between physicians, who set clinical performance requirements, and laboratory specialists, who assess feasible analytical performance, is essential. This dialogue is supported by simulation results, which assess how analytical performance affects a real-world dataset. The current analytical performance can then be interpreted as the probability that the clinical classification differs from the original study estimates that set the decision limits for the condition. If clinicians are satisfied with this level of uncertainty, the process concludes with mutual agreement. If not, and they seek to reduce uncertainty, the data simulation can be rerun: physicians specify the desired clinical performance, and the tool simulates various uncertainties to identify the necessary analytical performance. The laboratory specialist then evaluates whether achieving this performance is realistic given the conditions and timelines, and whether interim criteria are acceptable until the target performance is met. This process facilitates dialogue from "as good as it gets" to "as good as needed".

An example of using data simulation to study the impact of analytical performance on clinical diagnostic performance is illustrated by the performance of high-sensitivity cardiac troponin-T (hs-cTnT) in measuring myocardial damage, as outlined in the ESC-2020 guideline [11]. The above-mentioned simulation approach was used to assess the impact of current analytical performance (as determined by EQA) on the ability to correctly rule in or rule out myocardial infarction, using results from ~3,300 patients. For each patient result, different delta values were simulated to mimic clinical decisions according to the ESC 2020 guideline, and each result was simulated 10,000 times. The study showed that it was technically feasible to achieve a

clinical misclassification rate well below 1 %, as most laboratories demonstrated sufficient analytical performance. The minority that did not (about 20 %) needs to improve their analytical performance to meet the required clinical performance, as the simulation experiment showed that 2.2 % of patients were falsely ruled out when they ought to have been ruled in. Since the majority of laboratories used identical instruments and reagents that met the required analytical and clinical specifications, it became evident that improvement from "as good as it gets" to "as good as needed" is realistically feasible for this test in this clinical application.

Another example of the application of simulation studies to facilitate the dialogue between physicians and laboratorians is the performance of the immunochemical test for hemoglobin in feces (FIT), used for the referral of asymptomatic individuals to colonoscopy as part of national and regional colon cancer screening programs [15]. Once the referral rate in a particular screening program with a given incidence rate is established, data simulation enables the study of how analytical performance affects variation in the percentage of participants with a result above the threshold for colonoscopy referral. An excess of such variation does not only impact the opportunity of early diagnosis of treatable cancers and advanced adenomas but also has negative practical consequences for the system wherein such screening takes place. Over- or underuse of screening capacity has negative consequences for staff, clients, and reimbursement bodies, but also negatively impact the contingency for symptomatic patients that require colonoscopy. The Dutch programme for colon cancer screening uses an FIT threshold that yields an average referral rate of 4.5 % in the currently tested population [16]. Using historical FIT data in a simulation study, it was established that a ±0.5 % range around the ~5 % referral rate is desirable and achievable when the total analytical variation is less than 9.5 %, rounded to 10 % [17]. The Dutch programme has used this outcome not only to set the maximum total analytical variation for the external quality controls used to monitor the system, but also to divide that total budget among four main sources of that variation. Random within-day variation, between-analyser (and laboratory) variation, between-reagent-lot variation, and between-sample-buffer-and-tube-lot variation were all assigned an equal and feasible 5 % of that budget. Robust persistence of performance across all four sources of variation is monitored at a laboratory-transcending level, rendering the performance of the participating laboratories equivalent to that of one virtual analyser. To achieve that performance in terms of "as good as needed", in some cases reagent/calibrator lots had to be rejected.

In our current journey from focusing solely on analytical performance to emphasising clinical performance, we recognise that 1/a laboratory reports its initial measurement result, which may be the only one, and subsequent replicate measurements could vary, leading to different clinical classifications; 2/data simulations that incorporate relevant uncertainties help investigate how different estimates might impact outcomes. Such studies, in collaboration with clinicians, enhance the translation of results from clinical laboratories into better clinical decisions and, consequently, improved patient outcomes.

Model 1b to define analytical performance specifications for measurement uncertainty (Mauro Panteghini)

The role of measurements in laboratory medicine is to support medical decisions that lead to the best patient outcomes. Ensuring that the actual measurement uncertainty (MU) fits for the intended purpose is therefore essential in medical laboratories to provide high-quality information supporting clinical decisions [18]. The Milan model 1b for deriving APS relies on indirect methods that assess how the analytical performance of the laboratory test affects clinical classification and, ultimately, patient outcomes [2, 5, 6]. Model 1b-based APS establish limits for the maximum allowable measurement uncertainty of results from IVD MD used in conditions where: 1/decision thresholds and clinically acceptable misclassification rates have been agreed; 2/disease prevalences are known; and 3/the intended use of the IVD MD result has been agreed [18]. Therefore, Model 1b-derived APS are suitable when the measurand plays a central, well-defined role in decision-making for a specific disease, and the results are interpreted using established decision limits. This applies, for example, to fasting plasma glucose measurements and HbA1c in diagnosing and monitoring diabetes mellitus [19], to folate for diagnosing vitamin deficiency [20], and to cardiac troponins for diagnosing acute myocardial infarction [21–24]. Overall, there is a need for greater collaboration between laboratory and clinical professionals to define the key aspects required to derive and apply model 1b APS for MU in a practical, feasible, and least-biased manner [4, 25].

Simulation-derived information from patient results – APS Calculator/Simulator (Hikmet Can Çubukçu)

Simulation methods for establishing indirect outcome-based APS involve computer-generated replicates of actual measurement results under the simulated uncertainty conditions that existed when patient samples were measured. This allows for the comparison between original and simulated results, enabling the determination of APS that meet pre-established performance thresholds.

These simulation methods involve comparing the initial result with a repeated (second/simulated) measurement and determining the analytical performance characteristic that meets the preset agreement threshold. When setting APS for both imprecision and bias, analytical re-run and re-sampling simulations generally consider both characteristics simultaneously. By examining the effect of varying imprecision and bias on clinical classification, one can evaluate how current analytical performance may lead to potential misclassification in real-world practice [26].

On the other hand, acceptable bias is included within the long-term imprecision component of measurement uncertainty (MU) [27]. Conversely, and ideally unacceptable bias must be corrected as mandated by ISO 20914 [28]; it is therefore omitted from the calculations to prevent the inflation of MU estimates. Consequently, and consistent with traceability theory, MU is treated as the sole analytical performance characteristic in some simulations designed to set APS for MU.

The *APS Calculator* (web app, https://hikmetc-apscalculator.streamlit.app) was created to set APS for measurement uncertainty and to simulate an analytical re-run scenario. *APS Simulator* (https://github.com/hikmetc/APS_Simulator) includes all the features of the APS Calculator. Additionally, APS Simulator can determine APS for imprecision and bias through re-sampling and an analytical re-run scenario [29].

The *assumptions* of the simulation model are:

1. The measurement results used in the simulations are assumed to be non-biased
2. The preanalytical component of the variation in measurement results is not accounted for.
3. Clinical decision limits are ideally established to facilitate the assessment of diagnostic risks. In the absence of clinical decision limits, reference intervals, along with clinical symptoms, signs, and imaging results, may be utilised.
4. The input data set must come from the population of interest and should reflect the prevalence of the condition of interest.

While the APS Simulator provides valuable insights, significant challenges remain in the broader application of indirect outcome-based APS, including accounting for population characteristics, ensuring appropriate decision limits, and selecting appropriate performance thresholds (e.g., agreement). Addressing these limitations will be essential to enhancing the standardisation and robustness of APS determination.

Clinically acceptable analytical performance specifications (Emmi Rotgers)

Clinical professionals rely on established or intuitive decision limits when interpreting results from IVD MD in their practice. The intended use of the IVD MD results and the corresponding decision limits are usually established in conjunction with acceptable misclassification rates and disease prevalences.

Whether an IVD MD in a medical laboratory can provide a service suitable for its intended clinical purpose depends on the within-subject biological (CV_I) variation inherent to the patient population the clinical laboratory is serving, and the clinical traditions for diagnosis and monitoring practised by the clinical professionals using the services of the laboratory, as well as on the preanalytical (CV_{PRE}) and analytical variation (CV_A) associated with the actual measurand.

Clinically acceptable analytical performance specifications (CAAPS) [30, 31] for IVD MD are among the current methods for establishing Model 1b APS, and utilise the established reference change value concept [32–34]. Clinical decision limits or clinically significant differences (CD) based on clinical guidelines or expert opinions can be used to estimate the allowable diagnostic variation (CV_D) needed to verify clinically acceptable analytical performance specifications [30]. The CD is modelled as two decision limits with related normally distributed variations. The reference-change-value formula is then used to determine a budget for the allowable diagnostic variation. For simplicity, biological variation in health is used to model both variation in health and in disease. CAAPS serves as a clinical and end-user laboratory tool that can be adapted for various diagnostic or monitoring targets.

Model 1b APS, based on CAAPS modelling, is designed to be a clinically intuitive and practical tool for clinical

laboratories. However, using CAAPS requires careful consideration of the sensitivity and specificity associated with normal distribution modelling and the validity of the input data.

The simplicity of the current CAAPS version is an advantage. However, it still needs to be investigated whether a more complex CAAPS model, which, for example, includes clinically acceptable misclassification rates, the prevalence of the condition in the studied population, and biological variation in disease, can prove to be better than the current CAAPS model.

Guideline-driven approach to define Model 1b APS (Eric Kilpatrick)

The talk illustrated how guidelines can help establish analytical performance specifications (APSs), at least for some tests, using the example of blood ketone measurement. When establishing desirable APSs for blood ketones, some of the Milan options [2] are not appropriate. For example, direct outcome studies do not exist (Model 1a); the biological variability of ketones during diabetic ketoacidosis (DKA) is likely to be quite different compared to normality, thereby meaning Model 2 cannot be used; and applying the 'state of the art' (Model 3) is problematic as the fundamental issue is that we do not know if state of the art is clinically good enough or not. However, it might be possible to simulate how assay performance might impact clinical decisions. (Model 1b).

There is an emerging consensus around the blood ketone (beta-hydroxybutyrate, BOHB) concentration thresholds which should be used to categorise patients who are suspected as having DKA. There are 4 categories in general use, whose names vary slightly according to the guideline, but which generally indicate (in ascending concentrations) either normality, ketonaemia, impending DKA or probable DKA. In patients already known to have DKA, treatment success is also guided by a decrease in BOHB concentration of at least 0.5 mmol/h.

These two aspects of the BOHB test use (for diagnosis of DKA and for assessing the success of DKA treatment) were chosen to try to establish Model 1b clinical APSs. To help avoid the 'line problem' around a fixed BOHB decision limit with DKA categorisation, a clinical assumption was made to aim for analytical performance that ensured a low likelihood of patients being wrongly placed in a non-adjacent category. For example, in the 'impending DKA' category rather than normality, or vice versa, or the 'ketonaemia' category rather than 'probable DKA', or vice versa. Using this approach, a statistically sufficient within-lab CV_A of <21 % was, with >99 % certainty, sufficient to prevent such miscategorisation, although measurement instruments with bias could markedly reduce this CV_A; e.g., a 0.8 mmol/L bias reduces the acceptable CV_A to just 10 % [35, 36].

Using the 0.5 mmol/L/h desirable fall in BOHB as the basis for establishing the APS, another clinical assumption was made. This time, it was to ensure that an indicated drop in measurement of 0.5 mmol/L was not, in reality, a concentration that was unchanged or had indeed risen. Or that a reported unchanged concentration had not actually fallen by at least the recommended 0.5 mmol/L. Including the rounding errors introduced in measuring such small BOHB concentration changes, within-instrument APSs could be calculated, each depending on the initial BOHB concentration and the degree of certainty required e.g. 90 %, 95 %, 99 % [36].

This overall approach has several advantages, including that the derived APS directly relates to the clinical performance required for reliable guideline implementation. It is also straightforward to apply, using simple statistical tools. However, its use is limited to those tests where more than two diagnostic categories exist (HbA1c being another example), or where recommended concentration change limits in the same patient have been defined. The APS is also only applicable to the clinical situation described in the guideline; for example, the BOHB APS applies only to the use of the test for DKA, not, say, in subjects on a ketogenic diet for epilepsy.

Error grid analysis to define Model 1b APS (Katy Bell)

Error grid analysis is a well-established method for assessing performance using clinical decision limits [35, 37]. In common with many other Model 1b APS methods, it uses a simulation study to undertake 3 key steps: 1/calculation of '*true*' test values; 2/calculation of *measured* test values; and 3/calculation of the *impact* of discrepancies between 1 and 2 on outcomes [6]. Error grid analysis was first developed in the 1980s in a use case investigating the potential impact of measurement discrepancies between self-monitoring blood glucose devices and laboratory reference measurements on insulin dosing errors [38]. Using a scatter plot of reference vs. self-monitoring test results, the plot was divided into error-grid zones based on the potential clinical consequences of associated dosing errors [35, 38]. It was later expanded to evaluate the broader metrological and diagnostic properties of IVD MD [39, 40].

A highly cited application of this method by Boyd and Bruns investigated the potential impact of observed blood

glucose measurements on ICU nurses' adjustments of insulin dose delivered via infusion to critically unwell patients [41]. In the simulation, higher insulin doses were administered at higher measured glucose levels, according to predetermined rules for each patient, consistent with "sliding scales" for insulin infusion dosing used in clinical practice at the time [41]. They evaluated the potential impact of analytical performance on the ability to administer appropriate insulin doses, as indicated by the patient's true glucose concentration. Using Monte Carlo (re-sampling) simulation modelling, they generated glucose measurements with specified bias and imprecision and then estimated insulin dosage error rates for results affected by known imprecision and/or bias. A more recent application of the error grid approach investigated its use for Health Technology Assessment [42]. The scenario involved the use of faecal calprotectin as a triage test for specialist referral to investigate possible Inflammatory Bowel Disease (IBD), using data from 951 patients presenting to their General Practitioner in the UK [43, 44]. The impact of imprecision and bias (APS) on clinical diagnostic sensitivity and specificity was studied using replacement sampling (bootstrapping) and sampling from lognormal parametric distributions applied to the baseline data [43, 44].

The benefits of the error grid analysis method include: 1/visually appealing output; 2/transparent assumptions linking test results to clinical actions and outcomes; 3/simulation examples exploring different error values and distributions; 4/suitability for dosing decisions and therapeutic drug monitoring, where the link between test results and clinical actions is well validated. Its limitations include: 1/requiring clinician expert input or validation for error zone placement; 2/dependence on accurate data on the distribution of true measurand levels and disease status for simulation setup; 3/limited empirical data on the actual risk of adverse effects from incorrect dosing decisions or other clinical actions.

Placing Model 1b-derived APS in context, and re-assessing models based on changes in the clinical classification of patients (Graham Jones)

Constraints to outcomes of Milan Model 1b APS studies

Methods are being developed [30, 36] to establish APS based on patient classifications or clinical decisions [2], with the aim of enhancing patient outcomes. The Milan Model 1b APS [2] can address multiple assay performance criteria such as bias, imprecision, or total error, and can be applied in various contexts [45], including evaluating analytical performance in External Quality Assurance schemes, designing quality control protocols, and aiding method selection and verification. It also helps set acceptance criteria for sample stability and other pre-analytical studies [46]. Additionally, it guides producers of reference materials and providers of measurement services in meeting relevant standards [47]. However, the Milan Model 1b APS should not be interpreted or used in isolation, as factors such as biological variation, current technological standards, and assay performance in outcome studies set limits on what is meaningful or valid.

We propose that APS derived from Milan Model 1b should be assessed in the light of data from Milan Model 2 (biological variation) and Milan Model 3 (state-of-the-art). If, for example, an assay has imprecision (CV_A) which is small relative to within-subject biological variation (CV_I), any APS proposing even smaller imprecision, however derived, will have minimal effect on the final variation of the result used by the clinician. Analytical performance specifications that propose APS tighter than those in routine use may be difficult to apply in practice, and studies based on clinician opinion will be informed by the "state of the art", here meaning the assays with which the clinicians have gained their experience. In addition, proposing APS that require tighter performance than the assays used in clinical studies to determine clinical decision points cannot be fully supported. Consequently, if Model 1b produces APS that are narrower than these limits, such APS are either not clinically meaningful, technically feasible, or supported by available evidence. When establishing APS, it is important to consider information from other Milan APS Models and related factors [46].

Common Milan Model 1b APS simulation modelling systematically underestimates the inherent instability of patient classifications

Current Milan Model 1b APS simulation modelling commonly assesses the effects of changes in assay performance by changes in the clinical classification of subjects relative to an original data set [26, 29, 48–51]. These models treat the original dataset as the "true" classification and then simulate how bias or imprecision might alter these classifications. In reality, every clinical dataset already includes misclassification caused by within-subject biological variation (CV_I) and

analytical imprecision (CV_A). Consequently, when modelling APS, the relevant measure is not the absolute percentage of differently classified results but the change in alternative classification rate relative to the baseline instability of the classification in the original study.

Two primary issues are identified: *re-measurement instability*, which leads to different classifications upon repeated testing of the same samples, and *re-sampling instability*, in which re-studying the same population may also lead to reclassification of individuals near the decision threshold, despite consistent analytical performance. Hence, assuming "zero misclassification" in any published study used, for example, for setting decision limits for a condition, is unrealistic.

The proposed improved method consists of two interconnected parts: first, the assay performance should be modelled in relation to the assay used in the original study – CV_A changes are represented as increases or decreases compared to the study assay, with bias defined relative to the initial measurement process. Second, the baseline alternative classification rate must be determined with knowledge of the CV_I, CV_A, and the population density near the decision threshold. This baseline alternate classification rate can be described as the "line problem", where individuals with homeostatic set points close to the decision point may flip classification randomly with repeat testing.

The proposed alternative reframes APS determination relative to the underlying, unavoidable alternate classification rate that exists even in the absence of any change in assay performance. Implementing this approach on real datasets presents several challenges: they already include CV_I and CV_A; merely adding more variation is insufficient. A method is required to infer the underlying distribution of homeostatic set points, to integrate regression to the mean, and to establish acceptable limits for changes in alternate classification rates.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: All authors wrote and/or contributed to their parts of the manuscript. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Kenny D, Fraser CG, Petersen PH, Kallner A. Consensus agreement. Scand J Clin Lab Invest 1999;59:585.
2. 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.
3. Magnusson B, Örnemark U. Eurachem guide: the fitness for purpose of analytical methods – a laboratory guide to method validation and related topics. Eurachem; 2014. Available from: www.eurachem.org.
4. Horvath AR, Bossuyt PM, Sandberg S, John AS, Monaghan PJ, Verhagen-Kamerbeek WD, et al. Setting analytical performance specifications based on outcome studies – is it possible? Clin Chem Lab Med 2015;53:841–8.
5. Jones GRD, Bell KJL, Ceriotti F, Loh TP, Lord S, Sandberg S, et al. Applying the Milan models to setting analytical performance specifications – considering all the information. Clin Chem Lab Med 2024;62:1531–7.
6. 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.
7. Sandberg S, Zima T, Panteghini M. Analytical performance specifications – moving from models to practical recommendations. Clin Chem Lab Med 2024;62:1451–4.
8. Thue G, Sandberg S. Analytical performance specifications based on how clinicians use laboratory tests. Experiences from a post-analytical external quality assessment programme. Clin Chem Lab Med 2015;53:857–62.
9. Skeie S, Thue G, Sandberg S. Patient-derived quality specifications for instruments used in self-monitoring of blood glucose. Clin Chem 2001;47:67–73.
10. Skeie S, Thue G, Sandberg S. Interpretation of hemoglobin A(1c) (HbA(1c)) values among diabetic patients: implications for quality specifications for HbA(1c). Clin Chem 2001;47:1212–7.
11. van Schrojenstein Lantman M, Grobben R, van Herwaarden AE, van Berkel M, Schaap J, Thelen M. To rule-in, or not to falsely rule-out, that is the question: evaluation of hs-cTnT EQA performance in light of the ESC-2020 guideline. Clin Chem Lab Med 2024;62:1158–66.
12. Mueller C, Boeddinghaus J, Nestelberger T. Downstream consequences of implementing high-sensitivity cardiac troponin: why indication and education matter. J Am Coll Cardiol 2021;77:3180–3.
13. Kumar A, Patel DR, Menon V. Cardiac troponin testing: goodbye, 'troponinemia. Cleve Clin J Med 2020;87:483–4.
14. Devereaux PJ, Lamy A, Chan MTV, Allard RV, Lomivorotov VV, Landoni G, et al. High-sensitivity troponin I after cardiac surgery and 30-Day mortality. N Engl J Med 2022;386:827–36.
15. Young GP, Benton SC, Bresalier RS, Chiu HM, Dekker E, Fraser CG, et al. Fecal immunochemical test positivity thresholds: an international survey of population-based screening programs. Dig Dis Sci 2025;70:1637–45.
16. Toes-Zoutendijk E, Bonfrer JMG, Ramakers C, Thelen M, Spaander MCW, Dekker E, et al. Quality monitoring of a FIT-based colorectal cancer screening Program. Clin Chem 2019;65:419–26.

17. van Schrojenstein Lantman M, de Jong A, van der Groep P, Kemper E, de Groot M, Vossen R, et al. The Dutch Colon cancer screening programme as an example of adopting indirect-outcome-based analytical performance specifications using a virtual analyser concept. Clin Chem Lab Med 2026 Sep 11. https://doi.org/10.1515/cclm-2026-031 [Epub ahead of print].
18. Panteghini M, Krintus M. Establishing, evaluating and monitoring analytical quality in the traceability era. Crit Rev Clin Lab Sci 2025;62: 148–81.
19. Nielsen AA, Petersen PH, Green A, Christensen C, Christensen H, Brandslund 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.
20. Ferraro S, Lyon AW, Braga F, Panteghini M. Definition of analytical quality specifications for serum total folate measurements using a simulation outcome-based model. Clin Chem Lab Med 2020;58: e66–8.
21. Lyon AW, Kavsak PA, Lyon OA, Worster A, Lyon ME. Simulation models of misclassification error for single thresholds of high-sensitivity cardiac troponin I due to assay bias and imprecision. Clin Chem 2017; 63:585–92.
22. Krintus M, Panteghini M. Judging the clinical suitability of analytical performance of cardiac troponin assays. Clin Chem Lab Med 2023;61: 801–10.
23. Collet JP, Thiele H, Barbato E, Barthelemy O, Bauersachs J, Bhatt DL, et al. 2020 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Eur Heart J 2021;42:1289–367.
24. Li Z, Tew YY, Kavsak PA, Aakre KM, Jaffe AS, Apple FS, et al. Impact of high-sensitivity cardiac troponin I assay imprecision on the safety of a single-sample rule-out approach for myocardial infarction. Clin Chem Lab Med 2025;63:e59–62.
25. Horvath AR, Bell KJL, Ceriotti F, Jones GRD, Loh TP, Lord S, et al. Outcome-based analytical performance specifications: current status and future challenges. Clin Chem Lab Med 2024;62:1474–82.
26. Cubukcu 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.
27. Thelen MHM, van Schrojenstein Lantman M. When bias becomes part of imprecision: how to use analytical performance specifications to determine acceptability of Lot-Lot variation and other sources of possibly unacceptable bias. Clin Chem Lab Med 2024;62:1505–11.
28. ISO/TS 20914. Medical laboratories — practical guidance for the estimation of measurement uncertainty, 2019. Geneva, Switzerland: International Organization for Standardization; 2019.
29. Cubukcu HC, Vanstapel F, Thelen M, van Schrojenstein Lantman M, Bernabeu-Andreu FA, Mesko 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 2024;62: 597–607.
30. 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.
31. Rotgers E, Lamberg T, Pihlajamaa T, Pussinen C, Joutsi-Korhonen L, Kouri TT. Verifying measurements on Siemens Atellica(R) instruments using clinically acceptable analytical performance specifications. Scand J Clin Lab Invest 2023;83:408–16.
32. Barnett RN. Medical significance of laboratory results. Am J Clin Pathol 1968;50:671–6.
33. Fraser CG. Reference change values. Clin Chem Lab Med 2012;50: 807–12.
34. Fraser CG. Making better use of differences in serial laboratory results. Ann Clin Biochem 2012;49:1–3.
35. Parkes JL, Slatin SL, Pardo S, Ginsberg BH. A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. Diabetes Care 2000;23:1143–8.
36. Kilpatrick ES, Butler AE, Atkin SL, Sacks DB. Establishing pragmatic analytical performance specifications for blood beta-hydroxybutyrate testing. Clin Chem 2023;69:519–24.
37. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes Care 1987;10:622–8.
38. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes Care 1987;10:622–8.
39. 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.
40. Kovatchev BP, Wakeman CA, Breton MD, Kost GJ, Louie RF, Tran NK, et al. Computing the surveillance error grid analysis: procedure and examples. J Diabetes Sci Technol 2014;8:673–84.
41. Boyd JC, Bruns DE. Quality specifications for glucose meters: assessment by simulation modeling of errors in insulin dose. Clin Chem 2001;47:209–14.
42. Smith AF, Messenger M, Hall P, Hulme C. The role of measurement uncertainty in health technology assessments (HTAs) of in vitro tests. Pharmacoeconomics 2018;36:823–35.
43. Smith A, Shinkins B, Hulme C, Hall P, Messenger M. Methods to assess the impact of test measurement uncertainty on downstream clinical outcomes: a case study of faecal calprotectin (FC) for the diagnosis of inflammatory bowel disease (IBD). Clin Chim Acta 2019; 493:S353–4.
44. Smith A. Developing a framework for assessing the impact of test measurement uncertainty on clinical and health-economic outcomes. University of Leeds; 2020.
45. Jones GRD, Albarede S, Kesseler D, MacKenzie F, Mammen J, Pedersen M, et al. Analytical performance specifications for external quality assessment – definitions and descriptions. Clin Chem Lab Med 2017;55:949–55.
46. Jones GRD. Using analytical performance specifications in a medical laboratory. Clin Chem Lab Med 2024;62:1512–19.
47. 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;62:1490–6.
48. Loh TP, Markus C, Lim CY. Impact of analytical imprecision and bias on patient classification. Am J Clin Pathol 2024;161:4–8.
49. Petersen PH, de Verdier CH, Groth T, Fraser CG, Blaabjerg O, Horder M. The influence of analytical bias on diagnostic misclassifications. Clin Chim Acta 1997;260:189–206.
50. 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. https://doi.org/10.1038/s41598-017-14172-8.
51. 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.