Article:
Ping Loh, T, Smith, AF orcid.org/0000-0001-7709-1869, Bell, KJL et al. (6 more authors) (2021) Setting analytical performance specifications using HbA1c as a model measurand. Clinica Chimica Acta, 523. pp. 407-414. ISSN 0009-8981

https://doi.org/10.1016/j.cca.2021.10.016

Review

Title: Setting analytical performance specifications using HbA1c as a model measurand

Running title: Analytical performance specifications

Authors: Tze Ping Loh[1,*], Alison F Smith[2,3], Katy JL Bell[4], Sarah J Lord[5,6], Ferruccio Ceriotti[7], Graham Jones[8,9], Patrick Bossuyt[10], Sverre Sandberg[11,12,13], Andrea Rita Horvath[14] on behalf of the European Federation of Clinical Chemistry and Laboratory Medicine Task Group on Outcome-based Analytical Performance Specifications

[1]Department of Laboratory Medicine, National University Hospital, Singapore

[2]Test Evaluation Group, Academic Unit of Health Economics, University of Leeds, Leeds, UK

[3]NIHR Leeds In Vitro Diagnostic (IVD) Co-operative, Leeds, UK

[4]School of Public Health, The University of Sydney, New South Wales Australia.

[5]School of Medicine, University of Notre Dame, Darlinghurst, New South Wales, Australia

[6]NHMRC Clinical Trials Centre, The University of Sydney, Camperdown, New South Wales, Australia

[7]Clinical Laboratory, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy

[8]Department of Chemical Pathology, SydPath, St Vincent's Hospital, Sydney, New South Wales, Australia.

[9]University of New South Wales, Sydney, New South Wales, Australia

[10]Department of Epidemiology and Data Science, Amsterdam Public Health, Amsterdam University Medical Centers, the Netherlands

[11]Norwegian Organization for Quality Improvement of Laboratory Examinations (NOKLUS), Haraldsplass Deaconess Hospital, Bergen, Norway

[12]Norwegian Porphyria Centre, Department of Medical Biochemistry and Pharmacology, Haukeland University Hospital, Bergen, Norway

[13]Institute of Public Health and Primary Health Care, University of Bergen, Bergen, Norway;

[14]Department of Clinical Chemistry and Endocrinology, New South Wales Health Pathology, Prince of Wales Hospital, Sydney, Australia

*Corresponding author
Address: National University Hospital, 5 Lower Kent Ridge Road, Singapore 119074
Telephone: (+65) 67724345
Fax: (+65) 67771613
Email: tploh@hotmail.com

Keywords: analytical performance speficiation; bias; imprecision; external quality assurance; proficiency testing; quality goal; quality control

Non-standard abbreviations: APS, analytical performance specification; CVa, analytical coefficient of variation; CVi, within-subject biological variation; CVg, between-subject biological variation; RCV, reference change value

Abstract

Analytical performance specifications (APS) for measurands describe the minimum analytical quality requirements for their measurement. These APS are used to monitor and contain the systematic (trueness/ bias) and random errors (precision/ imprecision) of a laboratory measurement to ensure the results are “fit for purpose” in informing clinical decisions about managing a patient’s health condition. In this review, we highlighted the wide variation in the setting of APS, using different levels of evidence, as recommended by the Milan Consensus, and approaches. The setting of *a priori* defined outcome-based APS for HbA1c remains challenging. Promising indirect alternatives seek to link the clinical utility of HbA1c and APS by defining statistical confidence for interpreting the laboratory values, or through simulation of clinical performance at varying levels of analytical performance. APS defined based on biological variation estimates in healthy individuals using the current formulae are unachievable by nearly all routine laboratory methods for HbA1c testing. On the other hand, the APS employed in external quality assurance programs have been progressively tightened, and greatly facilitate the improved quality of HbA1c testing. Laboratories should select the APS that fits their intended clinical use and should document the data and rationale underpinning those selections. Where possible common APS should be adopted across a region or country to facilitate the movement of patients and patient data across health care facilities.

Introduction

With exponentially rising healthcare costs brought about by the demands of population growth and ageing coupled with rapid healthcare inflation, there is increasing emphasis on providing quality care at sustainable costs. The quality of diagnostic tests may have significant financial implications with people missing out on beneficial/ necessary treatment due to false negative results, or people undergoing further unnecessary investigations and treatment due to false positive results. On the other hand, requiring a test to perform better than that which is clinically required may add unnecessary costs by excluding less expensive methods that are still fit-for-purpose. At the healthcare systems level, biomarkers have been used as a convenient secondary indicator to assess the quality of care for certain diseases. This is most notable in chronic disease management such as diabetes and lipid disorders, where clear treatment goals, based on biomarker measurements, have been defined to assess adequacy of access to care and comparison of the quality of care [1]. Wide analytical bias and variation among laboratory methods can lead to erroneous conclusions about the quality of care, with far-reaching consequences that may lead to ill-informed policy decisions and suboptimal resource allocation [2]. This calls for analytical performance specifications for tests that meet the needs of clinicians who make decisions about the care of patients and policy makers who make decisions about health systems.

Analytical performance specifications (APS) for measurands describe the minimum analytical quality requirements for their measurement [3-5]. These APS are used to monitor and contain the systematic (trueness/ bias) and random errors (precision/ imprecision) of a laboratory measurement to ensure the results are “fit for purpose” in informing clinical decisions about managing a patient’s health condition. APS are often presented as discrete numerical or percentage values indicating a deviation from a defined target. They are used as assessment criteria by proficiency/ external quality assurance testing organizations and regulatory agencies, and play an important role in the accreditation and licensing of medical laboratories (e.g. the Clinical Laboratory Improvement Amendments 1988 in the US) and in the regulation and market approval of *in vitro* diagnostics (e.g. by the Food and Drug Administration in the US or Conformité Européenne marking in Europe). APS are the cornerstone of many clinical laboratory decisions

relating to quality, including the calculation of the sigma value of a laboratory measurement method and the setting of an internal quality control system (e.g. frequency, control limits, interpretative rules). Laboratories use APS as acceptance criteria when new methods or new lots of reagents are evaluated and when considering factors that may affect result such as hemolysis interference or analyte stability.

The process of setting APS for a specific measurand is a developing activity. In this paper we discuss the principles behind assigning APS using hemoglobin A1c (HbA1c) as a model measurand. An understanding of the principles in setting APS is key to appraising and applying them in routine laboratory practice. Hemoglobin A1c was selected as the model measurand as it has robust evidence base for different approaches in setting APS that can serve as illustrative examples.

Evidence-based analytical performance specifications

The 2015 Milan consensus proposed three models for setting APS for a measurand based on a hierarchy of evidence [4,5]. The highest level of evidence relates to health outcomes data (Model 1), followed by biological variation data (Model 2) and information on the analytical performance achieved by “state of the art” methods (Model 3). Model 1 can be approached in two ways, either directly with health outcome studies (Model 1a) or indirectly by modelling the influence of analytical performance on clinical decision making (Model 1b).

An ideal outcome study (Model 1a) would have an *a priori* objective of directly comparing outcomes in patients who are subjected to the *same* laboratory test/method of measurement, but with varying analytical performance [6]. The ‘outcomes’ of interest can be broadly categorized into health, operational, and economic. Health outcomes include patient outcomes such as morbidity, quality of life, and mortality. A positive difference in patient outcomes in this context provides direct evidence of the impact of analytical performance on clinical effectiveness. Operational and economic outcomes can also be taken into account when setting APS, provided that patient health outcomes are not unduly compromised. For example, HbA1c measured using a point of care

device with low bias in the clinic facilitates faster decisions on the adjustment of treatment, which can be better communicated to patients at their clinic visit and thus leads to increased adherence with treatment [7]. This raises the important concept of balancing priorities for both health outcomes and costs.

APS can also be defined using intermediate outcomes such as diagnostic accuracy and medical decision making, whereby the consequent management can be linked to evidence on patient outcomes to estimate clinical effectiveness (e.g. via decision analytic modelling; Model 1b)[8]. Cost-effectiveness studies, combining health and economic outcomes, (such as cost per quality adjusted life year) may also be conducted in this way.

In practice, there are significant challenges to executing studies that directly generate evidence on the impact of analytical performance on outcomes (Milan Model 1a). Notably, there may be ethical questions about conducting studies where the use of laboratory tests with differences in analytical variation and systematic bias may lead to different and potentially erroneous clinical decisions and thus harmful effects on health. Additionally, the operational challenges of maintaining the multitude of tests of varying analytical performance in the laboratory and ensuring that patients who are assigned to a certain analytical performance are consistently tested with this throughout the study period may not be trivial.

An alternative method is to examine the impact of analytical performance on health outcomes using quasi-experimental (pre-post intervention) study designs involving different laboratory methods of varying analytical performance [9]. Or, the measurand can be evaluated by different laboratory methods concurrently to examine the impact of different analytical performance on clinical decision making [10]. Simulation methods (i.e. Model 1b), based on the application of hypothetical bias and imprecision onto baseline test measurements, have also been employed in this context to overcome the challenges of conducting direct outcome studies [6].

Additionally, different laboratory assays for the same measurand commonly have varying analytical specificity (degree of interference/ cross-reactivity with non-target molecule), sensitivity (limit of detection) and upper reporting limits, which are dimensions of analytical performance specifications that can influence the interpretation and attribution of the observed outcome. For example, the presence of hemoglobin variants can variably interfere with HbA1c measurement [11,12]. This may lead to results being variably suppressed by different methods, if the interference is detected or not, or differences seen in monitoring a patient should their testing move between measurement systems. This may happen for example with a patient with persistent levels of hemoglobin F. Even within the same analytical method, different platforms can have different analytical quality, and this can affect diagnostic accuracy and subsequent patient management decisions [13]. Moreover, a change in laboratory method can also bring about a change in clinical workflow that may impact operational/ economic outcomes.

Based on the above, it is perhaps unsurprising that despite a strong mandate for laboratory medicine to base APS on outcome studies, they are exceedingly uncommon. We have little direct outcome data to support the objective setting of APS based on *a priori* determined clinical outcomes, and although simulation methods provide a potentially powerful mechanism for extracting APS, indirect outcome studies have yet to achieve wide-scale adoption in this context. Consequently, biological variation (Milan Model 2) and “state-of-the-art” (Milan Model 3) are currently the most used methods for setting APS in laboratory medicine.

Different uses of a measurand

A measurand has different uses in different clinical contexts and they may not always be apparent to the laboratory. At the first clinical encounter, a biomarker is often used to assist in the screening, clinical diagnosis or differential diagnosis of a patient [14]. The result of the biomarker is compared to a reference interval or a clinical decision limit in order to decide whether or not the patient has, or is likely to have the condition. Under such use, the clinical sensitivity (true positive rate) and specificity (true negative rate), and the consequent positive and negative predictive values (taking into account the disease prevalence in the target population) are of key importance. In such a

case, the APS should be set to minimize false positive and/or false negative rates to ensure diagnostic accuracy is optimal for the intended clinical use.

Stringent APS and a demand for harmonization or standardization between measurement methods are particularly important for measurands where a single threshold is used as a diagnostic criterion, such as HbA1c for diagnosis of diabetes mellitus [13]. For this setting, with regard to precision, the biological variation approach can be seen to be useful for indicating when the assay performance does not contribute materially to the total uncertainty of the result. By contrast, bias can lead to changes at the individual and population levels leading to changes in apparent disease prevalence.

In subsequent serial measurements of the same patient, the focus shifts from diagnosis to monitoring of the disease progression or response to treatment [15,16]. The reference change value (RCV) is the difference in sequential results that must be exceeded to be considered significant beyond the combined inherent biological and analytical variation in the two results. Mathematically, RCV (%) = $2^{0.5} \times Z \times [(CVa)^2 + (CVi)^2]^{0.5}$, where CVa is the within-laboratory analytical imprecision of the method, CVi is the within-subject biological variation and Z is the Z-value for a predefined probability. The result of the subsequent measurement is compared against the earlier results with the RCV to determine if a significant change that cannot be explained by analytical and biological variation with a certain probability has occurred [17]. However, it does not indicate anything concerning the probability that a true change has occurred or not [18]. An alternative view would be that, rather than dichotomising changes into significant/not-significant based on a pre-determined probability, a probability of a change being “real” can be calculated in the relevant clinical setting. A series of sequential measurements may also inform the general trend of the progression of the condition or its response to treatment. In all these considerations, the APS should be set such that the underlying trend of the serial measurements (the “signal”) is not obscured by analytical variation (the “noise”)[19]. Additionally, the systematic bias within and between analytical measurement methods needs to be understood to allow quantitative trending of

patient results to be identified across healthcare facilities that may be using different analytical methods.

Setting analytical performance specifications for HbA1c

Setting APS is a complex and multi-faceted exercise. We not only need to consider the evidence base, but also the potential use of the measurand in the context of the health system and the health care setting. To help illustrate these concepts, we present HbA1c as an example for setting APS (Table 1).

HbA1c is a biomarker that reflects the glycemic status of the last 30 - 60 days of an individual. It is used for monitoring the glycemic control of patients with diabetes using guideline-specific treatment targets. Since 2010, HbA1c has also been used as a diagnostic criterion for diabetes [20,21].

Analytical performance specifications for HbA1c based on Milan Model 1a

To date, there has been no study with a direct comparison of clinical outcomes in patients receiving HbA1c measurements using the same laboratory method with varying degrees of analytical performance.

Analytical performance specifications for HbA1c based on Milan Model 1b

The impact of analytical performance of tests on clinical outcomes can be investigated by statistical derivation, surveying clinicians, or by using modeling simulations.

Statistical derivation

To overcome the limitation of a lack of *a priori* clinical outcome studies, the APS for HbA1c have been statistically determined *a posteriori* using clinical outcomes measured in clinical trials.

The biochemical treatment target of diabetes mellitus is based on the risk of microvascular surrogate events (e.g. retinopathy detected on ophthalmologic screening or nephropathy detected on testing for albuminuria), first demonstrated in the Diabetes Control and Complications Trial and

UK Prospective Diabetes Study [22,23]. Different statistical approaches have been applied to trial data such as these, to derive acceptable within- and inter-laboratory analytical variation for HbA1c. (Table1)[24]. The original Diabetes Control and Complications Trial showed a significant difference in outcome with HbA1c of 7.0% and 8.0%. This sets a standard that these two values must be able to be separated (including biological variation). When HbA1c is used for monitoring disease progression, the RCV concept has been used to determine the analytical variation that allows the detection of a 0.5% National Glycohemoglobin Standardization Program (NGSP) unit change (~5.5 mmol/mol) in the absolute value of sequential HbA1c. This is commonly considered the target/ minimum response in treatment algorithms and clinical trials [24-26]. On the other hand, APS can also be determined if the required 95% confidence for a HbA1c treatment target (e.g. 7 ± 0.5% or 53 ± 5.5 mmol/mol) is defined. The APS determined under these two considerations are considerably different. Note also that both of these determinations only address the APS from a statistical point of view and do not directly link them with the clinical outcomes of the primary study. Based on these statistical methods of determining the APS, an inter-assay CVa of 5% (ideally <3%) has been recommended [27].

Physician survey

To link the APS of a biomarker to the clinical context within which it is used, clinical scenarios may be presented to clinicians to solicit their opinion of what is considered clinically significant. In a large international survey of this nature, which spanned six nations, clinicians were presented with a brief clinical scenario involving a patient with diabetes on a follow-up visit and asked the magnitude of change in HbA1c that is considered significant enough to warrant a change in clinical management [28]. The magnitude of change in HbA1c considered clinically significant by the responding clinicians was then used as a basis to determine the analytical variation, using the reference change value concept (Table 1). The scenario-based survey approach may improve the association of the APS with the clinical context. However, it may not be entirely reflective of routine clinical practice and is liable to clinician heuristics. It is also not directly related to clinically important health outcomes.

The range of responses produced very wide APS among the surveyed countries and were different for the expected direction of change [28]. For example, the range of CVa for the median RCV response from the 6 countries (at 80% probability) for positive direction of change was 3.5% to 11%, and 7.4% to 25% for negative direction of change. In other words, clinicians tolerate less analytical variation when HbA1c results increase than when results decrease, as increase in HbA1c triggers more clinical action than when results show a desirable downward trend in a treated diabetic patient. An alternative interpretation is that clinicians are aware that movement of results away from the population median are less likely than movements towards the median, due to the effect of regression to the mean. At 95% probability, it was not possible to calculate the CVa for the responses from several countries (Table 1). In part, this was due to the relatively high CVi estimate of 4% used in the study, which invalidated the calculation of the APS using the RCV method. An important limitation of this method is that the survey responses may have depended on the magnitude of analytical error in the laboratory tests clinicians were familiar with [29].

Statistical simulation modelling

Another approach to overcome the inherent challenges with *a priori* outcome studies is the use of statistical simulation modelling. Several approaches have been undertaken to examine the impact of analytical performance, in terms of imprecision and bias, on the clinical performance of HbA1c.

Under this approach, the distribution of HbA1c in a reference population is first evaluated, usually from a cross-sectional population survey. This baseline is then considered the 'true' value of the individual or population distribution [13,30,31]. Alternatively, the normal or log-normal distribution of published data has been adopted [32-34]. In reality, this distribution will contain confounding analytical variability of the laboratory method, which is often ignored. Following this, different degrees of analytical variation and bias can be artificially introduced to the distribution, to generate HbA1c values that incorporate these additional analytical errors. The diagnostic threshold of HbA1c for diabetes is then applied to the original HbA1c values as well as to those incorporating the analytical errors. Once this basic model is set up, diagnostic performance parameters can be examined, such as false positive and negative rates [30,34], or, alternatively, clinical sensitivity and

clinical specificity [13], for varying levels of analytical variation and bias. It should be noted that the latest definition of pre-diabetes uses a diagnostic threshold that may lie close to the average (mean, median) of HbA1c measurements in the general population. Consequently, the classification error can be very sensitive to changes in analytical variation and bias, which may result in both under- and over-diagnosis [13,30,34].

The use of real-world data allows for a numerical assessment of the effect of analytical bias on patient misclassification. In the following example, data was collected in an Australian laboratory over 3 years from patients with a single HbA1c measurement, assuming those were considered for diagnosis of diabetes rather than for monitoring of known diabetes (Figure 1). At a decision point of 5.5 % NGSP units, a positive bias of 0.2 % NGSP units (~3.6%), decreased the percentage flagged from 45% to 31%, with a similar negative bias increasing the flagged rate to 62%. By contrast biases of 0.3 % NGSP (4.6%) at the diabetes threshold of 6.5% changed the flagging rate from 12% to either 10% or 15%. In this model the use of the same percentage (or absolute) APS for HbA1c has markedly different numerical effects at different concentrations. This must be balanced against the critical importance of a diagnosis of diabetes against the less critical assessment of pre-diabetes. This is an argument for variable APS based on the value of the measurement result.

The advantage of this modelling approach is that it links analytical performance to clinical performance of HbA1c that is relevant to clinical decisions. It does not provide a definitive boundary for the APS, since the relationship between analytical performance and clinical performance is continuous. For example, increasing positive bias in HbA1c measurement will lead to exponentially increasing false positive rate for diabetes diagnosis [30]. In one study, this approach was used to examine the prognostic ability of HbA1c in detecting retinopathy [33]. Nevertheless, the clinical consequences of false negative and false positive diagnosis remain under-explored and present an important area of research. Further investigation could focus on subjects with values near the clinical decision points. For example, the misdiagnosis of a patient with a true HbA1c within 0.3 NGSP % units (above or below) 6.5% may be more common, but of

less consequence, than the misclassification of patients with true HbA1c results that are clearly in diabetes or normal ends of the population distribution. Of interest, one group has examined the impact of analytical performance on the probability of producing unreliable results, defined as a result exceeding a desirable APS [35]. Such studies could be redesigned to derive outcome-based APS.

It is also important to model the described use of a test as closely as possible. For example in Australia a repeat HbA1c >6.5% (48 mmol/mol) is required for the diagnosis of diabetes. While assay bias will have a clear effect on the diagnostic rate, the effect of imprecision is less clear. As only positive first results are repeated, the larger imprecision on the second measurement will also counteract the effect of over-classification on the first.

Depending on the statistical criteria chosen, the APS obtained using the indirect modelling approach may be relatively close to those derived *a posteriori* from clinical outcome studies, as described above (see Table 1), but this has not yet been formally evaluated. Further, the impact of HbA1c APSs on operational and economic outcomes remains under-investigated. In the general climate of resource consciousness, it may be timely to include an independent parameter, such as the laboratory cost of achieving different APS, to objectively select the most cost-effective discrete APS that optimally balances costs and health outcomes. For example, improved precision of results may be possible by analysis in duplicate, or bias improved by distribution of commutable samples to use as calibrators, however the cost implications cannot be ignored in this scenario.

More sophisticated decision modelling approaches can also be set up to examine the population from a health systems perspective. For example, the entire disease journey of a patient from time of diagnosis, treatment, development of complication, death, or other end outcomes can be mapped out. The corresponding probabilities of different permutations of the journey and the associated costs can be assigned, based on published data. A hypothetical group of patients that are representative of a reference population are then introduced into the model and allowed to go through the disease journey. The end outcomes of the patients are then recorded and evaluated.

Such an approach is commonly used to evaluate the cost-effectiveness of a healthcare intervention [36]. The impact of HbA1c measurement with different APS on healthcare costs and patient outcomes can be examined as part of the sensitivity testing of such models. Nonetheless, when a range of APS meets a predefined outcome (e.g. healthcare costs) in the sensitivity testing, a discrete APS may need to be selected subjectively.

Analytical performance specifications for HbA1c based on Milan Model 2

CVi is the day-to-day variation around a homeostatic set-point of a measurand. Between-subject biological variation (CVg) is the variation in the set-point among different individuals. There are well described methods to derive APS based on biological variation data [19]. The general principle behind these methods is to contain the analytical variation ('noise') relative to the biological 'signal' when determining the RCV, or to minimize the shift in population values relative to the reference intervals. Commonly, the APS are described as minimum, desirable, and optimum according to the degree of additional variability contributed by the analytical variation.

While this concept is elegant in its simplicity, its application may not be straightforward. It does not take into account the uncertainty of point estimate of the CVi which can vary from measurand to measurand. The conventional theories of biological variation are used in the above discussion and summary of results in Table 1 for consistency across publications from different periods. However, it should be noted that a more contemporary approach to biological variation and the derivation of APS has been proposed [37].

Biological variation data for HbA1c have been reported in multiple studies. The European Federation of Clinical Chemistry and Laboratory Medicine Task Group maintains a Biological Variation Database and APS for different measurands that are constantly updated to take into account the latest available evidence (https://biologicalvariation.eu/). It should be noted that these estimates contain considerable between-study variation.

The Task Group reviewed the literature and performed a meta-analysis of the HbA1c CVi and CVg in healthy subjects [38]. The CVg ranged widely, whilst the CVi estimates were more consistent (Table 1). The biological variation was found to be wider in subjects with diabetes even when well controlled [39]. Moreover, the biological variation estimates seemed to increase with higher concentration. The biological variation is even wider in children [40]. Care should be exercised when selecting the optimal biological variation estimates to derive the APS as they differ widely between populations and between different study designs. The CVi of HbA1c in healthy subjects is very small and thus the resulting desirable APS can be very difficult to achieve in routine laboratory practice (Table 1).

A weakness of the concept of CVg, i.e. the distribution of a measurand in healthy subjects, is that assessment of bias is related only to the misclassification of this population. In reality, subjects with the relevant disease may also be misclassified. The use of a population in which the test will be applied (see Australian example above), allows incorporation of both groups in the assessment. For the purposes of HbA1c, where clinical decision points rather than population reference intervals are used, the concept of CVg of a healthy population could be considered not relevant for the setting of APS.

An additional factor to the apparent need for very tight assay precision can be seen in the reporting intervals used for HbA1c. In NGSP units, the difference between 6.5 and 6.4% (NGSP units) is 1.5%, and between 48 and 47 mmol/mol is 2.1%. If an assay reaches a CVa of 1.2% (i.e. equal to the stated CVi), this will increase the variation in the results by a factor of approximately 1.4. On many occasions this will be within the same reporting interval and be unseen by the clinician [41] making the effort put into achieving this performance of limited clinical value

The biological variation approach to setting APS has two specific advantages. Firstly, if the goals are met, it can be stated with some certainty that the assay variation is small compared to the biological variation and will unlikely affect clinical decision-making; and secondly that the effect of

406 the analytical performance can be quantitated, providing information on the additional variation
407 seen in the final result.
408
409 ***Analytical performance specifications for HbA1c based on Milan Model 3***
410 The use of state-of-the-art to set APS is a pragmatic approach. It considers the capabilities of
411 current laboratory methods and sets the APS such that most of the laboratories can meet the
412 specification; for example the 95th percentile of peer laboratory performance [42]. This approach is
413 commonly used in external quality assurance/ proficiency testing programs and may carry
414 regulatory implications to ensure most (if not all) clinical laboratories can meet the requirements.
415 However, this approach in setting APS completely delinks the performance specifications from the
416 clinical utility of the measurand or the relationship between the analytical noise and biological
417 signal (variation). Indeed, this approach runs the risk of accepting laboratory methods that are not
418 fit for clinical purpose to be used in patient care [43]. Moreover, the definition of 'state-of-the-art'
419 can differ from country to country, in different publications or in different use, and requires
420 harmonization of analytical methods to ensure consistent clinical application. There is of course a
421 link between state of the art and physician survey for model 1b as described above, as clinicians
422 will have formed their opinion based on their experience with the quality of available
423 methodologies.
424
425 The adoption of a lenient state-of-the-art for setting APS reduces the incentive for the industry and
426 for laboratories to improve their analytical methodology and may lead to tolerance of poor
427 analytical performance. For HbA1c, a state-of-the-art APS for total error of up to 18%, which is
428 considerably higher than other approaches, has been adopted by a national proficiency testing
429 program, which is considering a revision [44]. The National Glycohemoglobin Standardization
430 Program has progressively tightened the APS for HbA1c over two decades, from an allowable total
431 error of ±12% in 2008 to ±6% currently, which led to significant improvement in test performance
432 [24]. The program is considered a great success story of assay standardization [44]. At the same
433 time, the use of sigma metrics as an assessment criterion has been evaluated and a clinical risk-
434 based approach has been proposed [45]. Under that approach, a laboratory achieving high sigma

performance is associated with higher passing rate in the proficiency testing program, although the optimal target sigma value is still debated [46]. Amidst the concerted effort by the industry and laboratory community to improve the standards for assessment, the Clinical Laboratory Improvement Amendments have recently proposed to widen the total error acceptance limit for HbA1c proficiency testing from ±6% to ±10%, which is widely considered as a major step backward that may put patients at risk [25].

Future: computer algorithm-driven clinical decision support

As healthcare progresses towards increasing adoption of computer algorithm-driven clinical decision support, such as rule-based algorithms, machine learning algorithms and other forms of artificial intelligence, the need for understanding the impact of analytical performance on these novel techniques is greater than ever [47]. Several groups have demonstrated feasibility of computer algorithm-driven clinical decision support tools for managing diabetes that use HbA1c as the key biomarker. The APS required to sustain the performance of these algorithms should be i) examined during the initial set-up as part of the sensitivity testing; ii) adopted by the clinical laboratory to ensure the robustness of the clinical decision support system; and iii) maintained using a quality management system.

Setting analytical performance specifications – the process

When an APS is set or adopted by an organization, the use of the above models allows a structured approach to selecting the most appropriate model taking into account the quality of the evidence, the clinical use of the assay (e.g. diagnosis or monitoring), the likely clinical setting (e.g. point of care) and available technologies, and the consequences of test results in terms of patient management. While a single model may be selected as the primary approach to setting APS, awareness of the available data and assessment of its quality for the other models allows a full awareness of the effect of selected APS against the other frameworks.

To adapt the Guide to the Expression of Uncertainty in Measurement [48]: While the Milan models provide a framework for assessing APS, they cannot substitute for critical thinking, intellectual honesty and professional skill. The evaluation of assay performance requirements is neither a

routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand, the measurement and the clinical use and consequence of the result. The quality and utility of the APS therefore ultimately depend on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value.

Conclusion

There is wide variation in the setting of APS, using different levels of evidence and approaches. The setting of *a priori* defined outcome-based APS for HbA1c remains challenging owing to complexities associated with conducting direct outcome studies. Promising indirect alternatives seek to link the clinical utility of HbA1c and APS by defining statistical confidence for interpreting the laboratory values, or through simulation of clinical performance at varying levels of analytical performance. APS defined based on biological variation estimates in healthy individuals using the current formulae are unachievable by nearly all routine laboratory methods for HbA1c testing. On the other hand, the APS employed in external quality assurance programs have been progressively tightened, and greatly facilitate the improved quality of HbA1c testing. Health economic modelling to estimate cost-effectiveness of various APSs for HbA1c may provide important information for laboratory and health policy decisions. Laboratories should select the APS that fits their intended clinical use and should document the data and rationale underpinning those selections. Where possible common APS should be adopted across a region or country to facilitate the movement of patients and patient data across health care facilities. This need for commonality indicates the vital role of professional organizations in achieving common practices.

Acknowledgement

The authors thank Eric Kilpatrick for his helpful comments that improved the manuscript.

Competing Interest

The authors have none to declare.

References

1. American Diabetes Association. Glycemic Targets: Standards of Medical Care in Diabetes-2021. Diabetes Care. 2021;44:S73-S84.

2. Carlsen S, Thue G, Cooper JG, Røraas T, Gøransson LG, Løvaas K, et al. Benchmarking by HbA1c in a national diabetes quality register--does measurement bias matter? Clin Chem Lab Med 2015;53:1433-1439.

3. Kallner A, McQueen M, Heuck C. The Stockholm Consensus Conference on quality specifications in laboratory medicine, 25-26 April 1999. Scand J Clin Lab Invest 1999;59:475-476.

4. 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;53:833-835.

5. 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;55:189-194.

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;65:1363-1374.

7. Tollånes MC, Jenum AK, Berg TJ, Løvaas KF, Cooper JG, Sandberg S. Availability and analytical quality of hemoglobin A1c point-of-care testing in general practitioners' offices are associated with better glycemic control in type 2 diabetes. Clin Chem Lab Med 2020;58:1349-1356.

8. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29:E1-E12

9. Wabe N, Li L, Lindeman R, Yimsung R, Dahm MR, McLennan S, et al. Impact of Rapid Molecular Diagnostic Testing of Respiratory Viruses on Outcomes of Adults Hospitalized with Respiratory Illness: a Multicenter Quasi-experimental Study. J Clin Microbiol 2019;57:e01727-18.

10. Sandoval Y, Smith SW, Thordsen SE, Bruen CA, Carlson MD, Dodd KW, et al. Diagnostic Performance of High Sensitivity Compared with Contemporary Cardiac Troponin I for the Diagnosis of Acute Myocardial Infarction. Clin Chem 2017;63:1594-1604.
11. Little RR, Rohlfing CL, Hanson S, Connolly S, Higgins T, Weykamp CW, et al. Effects of hemoglobin (Hb) E and HbD traits on measurements of glycated Hb (HbA1c) by 23 methods. Clin Chem 2008;54:1277-1282.
12. Loh TP, Cheng WL, Kao SL, Thai AC, Sethi SK. Effects of haemoglobin E traits on HbA1c measurement by two cation-exchange HPLC and two immunoturbidimetric methods. Pathology 2014;46:265-266.
13. 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; 7:13721.
14. Jones G, Barker A. Reference Intervals. Clin Biochem Rev 2008; 29: S93–S97.
15. Glasziou P, Irwig L, Mant D. Monitoring in chronic disease: a rational approach. BMJ 2005;330:644-648.
16. Bell KJ, Glasziou PP, Hayen A, Irwig L. Criteria for monitoring tests were described: validity, responsiveness, detectability of long-term change, and practicality. J Clin Epidemiol 2014;67:152-159.
17. Fraser CG. Changes in serial results. In: Biological variation: from principles to practice. Washington: AACC Press, 2001 67-90.
18. Petersen PH, Sandberg S, Iglesias N, Sölétormos G, Aarsand AK, Brandslund I, et al. 'Likelihood-ratio' and 'odds' applied to monitoring of patients as a supplement to 'reference change value' (RCV). Clin Chem Lab Med 2008;46:157-164.
19. Fraser CG, Hyltoft Petersen P, Libeer JC, Ricos C. Proposals for setting generally applicable quality goals solely based on biology. Ann Clin Biochem 1997;34:8-12.
20. American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021. Diabetes Care. 2021;44:S15-S33.

21. World Health Organization. Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus. https://www.who.int/diabetes/publications/report-hba1c_2011.pdf?ua=1 (Accessed Feb 2021)
22. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, Lachin J, Cleary P, Crofford O, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993;329:977-986.
23. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet 1998;352:837-853.
24. Little RR, Rohlfing CL, Sacks DB; 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;57:205-214.
25. Klonoff DC, Aron D, Cohen RM, Home P, John WG, Little RR, et al. The Need for Accuracy in Hemoglobin A1c Proficiency Testing: Why the Proposed CLIA Rule of 2019 Is a Step Backward. J Diabetes Sci Technol 2019;13:424-427.
26. Weykamp CW, Mosca A, Gillery P, Panteghini M. The analytical goals for hemoglobin A(1c) measurement in IFCC units and National Glycohemoglobin Standardization Program Units are different. Clin Chem 2011;57:1204-1206.
27. Sacks DB, Arnold M, Bakris GL, Bruns DE, Horvath AR, Kirkman MS, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2011;57:e1-e47.
28. Skeie S, Perich C, Ricos C, Araczki A, Horvath AR, Oosterhuis WP, et al. Postanalytical external quality assessment of blood glucose and hemoglobin A1c: an international survey. Clin Chem 2005;51:1145-1153.
29. 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;53:857-862.

30. 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;52:1069-1077.
31. Åsberg A, Odsæter 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;53:1459-1464.
32. Larsen ML, Fraser CG, Petersen PH. A comparison of analytical goals for haemoglobin A1c assays derived using different strategies. Ann Clin Biochem 1991;28:272-278.
33. Petersen PH, Jørgensen LG, Brandslund I, De Fine Olivarius N, Stahl M. Consequences of bias and imprecision in measurements of glucose and hba1c for the diagnosis and prognosis of diabetes mellitus. Scand J Clin Lab Invest Suppl 2005;240:51-60.
34. Petersen PH, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using Guideline-Driven Medical Decision Limits Clin Chim Acta 2014;430:1-8.
35. Woodworth A, Korpi-Steiner N, Miller JJ, Rao LV, Yundt-Pacheco J, Kuchipudi L, et al. Utilization of assay performance characteristics to estimate hemoglobin A1c result reliability. Clin Chem 2014;60:1073-1079.
36. Yang Y, Abel L, Buchanan J, Fanshawe T, Shinkins B. Use of Decision Modelling in Economic Evaluations of Diagnostic Tests: An Appraisal and Review of Health Technology Assessments in the UK. Pharmacoecon Open 2019;3:281-291.
37. Oosterhuis WP, Sandberg S. Proposal for the modification of the conventional model for establishing performance specifications. Clin Chem Lab Med 2015;53:925-937.
38. González-Lao E, Corte Z, Simón M, Ricós C, Coskun A, Braga F, et al. Systematic review of the biological variation data for diabetes related analytes. Clin Chim Acta 2019;488:61-67.
39. Carlsen S, Petersen PH, Skeie S, Skadberg Ø, Sandberg S. Within-subject biological variation of glucose and HbA(1c) in healthy persons and in type 1 diabetes patients. Clin Chem Lab Med 2011;49:1501-1507.
40. Desmeules P, Cousineau J, Allard P. Biological variation of glycated haemoglobin in a paediatric population and its application to calculation of significant change between results. Ann Clin Biochem 2010;47:35-8.

41. Jones GR. Effect of the reporting-interval size on critical difference estimation: beyond "2.77". Clin Chem 2006;52:880-5.
42. Ricós C, Ramón F, Salas A, Buño A, Calafell R, Morancho J, et al. Minimum analytical quality specifications of inter-laboratory comparisons: agreement among Spanish EQAP organizers. Clin Chem Lab Med 2011;50:455-461.
43. Heinemann L, Kaiser P, Freckmann G, Grote-Koska D, Kerner W, Landgraf R, et al. Higher HbA1c Measurement Quality Standards are Needed for Follow-Up and Diagnosis: Experience and Analyses from Germany. Horm Metab Res 2018;50:728-734.
44. Little RR, Rohlfing C, Sacks DB. The National Glycohemoglobin Standardization Program: Over 20 Years of Improving Hemoglobin A1c Measurement. Clin Chem 2019;65:839-848.
45. Weykamp C, John G, Gillery P, English E, Ji L, Lenters-Westra E, et al. Investigation of 2 models to set and evaluate quality targets for hb a1c: biological variation and sigma-metrics. Clin Chem 2015;61:752-759.
46. Little RR, Rohlfing CL. Assessing quality from an accuracy-based HbA1c proficiency survey. Clin Chem Lab Med 2016;54:e75-76.
47. Badrick T, Banfi G, Bietenbeck A, Cervinski MA, Loh TP, Sikaris K. Machine Learning for Clinical Chemists. Clin Chem 2019;65:1350-1356.
48. The International Bureau of Weights and Measures. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf (Accessed Feb 2021)
49. Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. Diabetes Care 2009;32:193–203.
50. Centre for Clinical Practice at NICE (UK). Type 2 Diabetes: Newer Agents for Blood Glucose Control in Type 2 Diabetes. London: National Institute for Health and Clinical Excellence (UK); 2009 May.

632 51. Rohlfing CL, Parvin CA, Sacks DB, Little RR; NGSP Steering Committee. Comparing analytic
633 performance criteria: evaluation of HbA1c certification criteria as an example. Clin Chim Acta
634 2014;433:259-263.

Table 1. Published examples illustrating different models for setting analytical performance specifications for HbA1c (data refer to measurements in NGSP% HbA1c units, unless otherwise stated). CVi = within-subject biological variation, CVg = between-subject biological variation, CVa = within-laboratory analytical variation, APS = analytical performance specification, EQA = external quality assurance program.

Evidence base	Model	Statistical approach	References	APS
Relationship between HbA1c and risk of progression to microvascular complication. Intensive Glycemic control (reflected by HbA1c) reduces microvascular complications	1b (Statistical derivation)	To interpret a treatment target of 7 ± 0.5% (53 ± 5.5 mmol/mol) with 95% confidence.	DCCT 1993 [22], UKPDS 1998 [23], Little 2011 [24]	CVa = 3.5% Assumed CVi = 1% (healthy subjects), bias = 0%
A reduction of 0.5% (5.5 mmol/mol) HbA1c is the evaluation target in treatment algorithms and commonly used in clinical trials.	1b (Statistical derivation)	HbA1c reduction by 0.5% NGSP units in sequential measurement is taken as critical difference. Applying the reference change value concept, RCV (%) = $2^{0.5} \times 1.96 \times [(CVa)^2 + (CVi)^2]^{0.5}$ to derived CVa. A Z-value of 1.96 represents a two-tailed 95% probability.	Little 2011 [24], Nathan 2009 [49], NICE 2009 [50]	CVa = 2% Assumed CVi = 1% (healthy subjects)
Physicians provided separate magnitude of change in HbA1c to be considered clinically significant change for clinical scenarios involving a positive and negative direction of change.	1b (Physician survey)	RCV used to calculate CVa based on the physician responses. One-tailed Z-values of 80% and 95% probabilities were used as the direction of change was specified in the survey.	Skeie 2005 [28]	Range of CVa for the median RCV response from the 6 countries (at 80% probability) for negative direction of change = 7.4% - 25% Range of CVa for the median RCV response from the 6 countries (at 80% probability) for positive direction of change = 3.5% - 11% Range of CVa for the median RCV response from the 6 countries (at 95% probability) for negative direction of change = 2.2% - 9.1% (not possible to calculate for 1 country) CVa for the median RCV response from the 6 countries (at 80% probability) for positive direction of change = 1.5% (not possible to calculate for all but 1 country) Assumed CVi = 4%, bias = 0%
Empirical national health survey data were considered ‘true’ values. CVi, CVa and bias were incorporated by simulation. The original and simulated results were	1b (Simulation)	Misclassification rates for different magnitudes of CVa (2% and 3.5%) and bias (-3% and + 2.5%) were assessed.	Chai 2017 [13]	APS not discreetly defined. Misclassification rates for different diagnostic thresholds demonstrated. Assumed CVi = 1.8% (healthy subjects)

classified according to diagnostic thresholds.				
Empirical HbA1c results from patients attending outpatient clinic of a single centre, with and without diabetes (defined by oral glucose tolerance test according to WHO criteria) were subjected to ± 0.6% HbA1c bias at 0.1% steps.	1b (Simulation)	The likelihood ratio of diagnosing diabetes with and without bias were compared.	Åsberg 2015 [31]	APS not discreetly defined. Change in likelihood ratio of diagnosing diabetes at different degrees of bias demonstrated. CVa not explored.
HbA1c in healthy reference populations that were represented by simulated normal/ log-normal frequency and cumulative probability distributions based on published data. CVa and bias were incorporated by simulation.	1b (Simulation)	A range of CVa and bias were assessed in terms of false positive rates above the diagnostic threshold.	Petersen 2014 [34]	At 1% false positive rate, assuming bias = 0%, CVa = 5% At 1% false positive rate, assuming CVa = 0%, bias = 3% Assumed CVi = 1.2%
A series of 'biological set-points' were assessed along the measurement range, with normal distributions fitted around each value to incorporate biological variability. Imprecision and bias was applied to each "true" set point by simulation.	1b (Simulation)	Probabilities of observing particular values for a given set point of HbA1c for the prognostic assessment of the risk of retinopathy were assessed	Petersen 2005 [33]	At 95% probability, assuming bias = 0%, CVa = 5% At 95% probability, assuming CVa = 0%, bias = 0.5%HbA1c Assumed CVi = 1.9%
Empirical data from a biobank from subjects without prior diagnosis of diabetes were extracted. CVa and bias were incorporated by simulation. Diagnosis of diabetes is based on fasting plasma glucose and HbA1c.	1b (Simulation)	Undiagnosed rate of diabetes for varying degree of CVa and bias were assessed.	Nielsen 2014 [30]	APS not discreetly defined. Undiagnosed cases of diabetes increased exponentially with increasing CVa and bias.
European Federation of Clinical Chemistry and Laboratory Medicine Working Group on Biological Variation and Task Group for the Biological Variation Database	2 (biological variation)	Meta-analysis of 13 studies on HbA1c biological variation derived from healthy individuals	González-Lao 2019 [30]	CVi = 1.3%, CVg = 5.0% CVa: minimum = 1.0%, desirable 0.7%, optimum 0.3% Bias: minimum 1.9%, desirable 1.3%, optimum 0.6% Total error: minimum 3.5%, desirable 2.4%, optimum 1.2%

Fifteen subjects with type 1 diabetes with <±1% HbA1c variation over 18 months and no changes in basal insulin dose over last 2 months and stable body weight. Samples collected weekly for 13 weeks	2 (biological variation)	CVi and CVg were estimated separately using an ANOVA model (balanced two-fold nested random model).	Carlsen 2011 [39]	CVi = 1.7%, CVg = 8.2% CVa: minimum = 1.3%, desirable 0.9%, optimum 0.4% Bias: minimum 3.1%, desirable 2.1%, optimum 1.0% Total error: minimum 5.2%, desirable 3.6%, optimum 1.7%
Thirty-six children with cystic fibrosis but do not have diabetes were included. At least 5 measurements taken over 5 years. Median age 14 years (range: 5-18 years)	2 (biological variation)	CVi and CVg were estimated separately using a nested ANOVA	Desmeules 2010 [40]	CVi = 4.8%, CVg = 12.8% CVa: minimum = 3.6%, desirable 2.4%, optimum 1.2% Bias: minimum 5.1%, desirable 3.4%, optimum 1.7% Total error: minimum 11.0%, desirable 7.4%, optimum 3.7%
College of American Pathologists	3 (state-of-the-art)	Assessment of participant results against target value set by higher metrology order techniques using APS.	Little 2019 [44]	Total error: ±5%
German Rili-BAEK. Previous proficiency testing material had commutability issue, necessitating a wider APS	3 (state-of-the-art)	Assessment of participant results against target value set by higher metrology order techniques using APS.	Heinemann 2018 [43]	At EQA APS = ± 18%, passing rate = 93% At EQA APS = ± 8%, passing rate = 83 %
The National Glycohemoglobin Standardization Program	3 (state-of-the-art	Assessment of participant results against target value set by higher metrology order techniques using APS. Contours plots of constant probability (0.95, 0.99 and 0.999) were derived from the computed probabilities of passing a given criterion over the grid of relative bias and CV combinations evaluated.	Little 2019 [44], Rohlfing 2014 [51]	Manufacturer and Level II laboratory: 36 of 40 results within ±5% Level I laboratory: 37 of 40 results within ±5%
Clinical Laboratory Improvement Amendments Rule of 2019 (proposed)	3 (state-of-the-art)	Assessment of participant results against target values using APS.	Klonoff 2019 [25]	± 10%

Figure legend

Figure 1. Selected population representing those being assessed for diagnosis of diabetes (as described in the text). The fraction of the population re-classified with +/- 0.2 units at 5.5% (NGSP) and +/- 0.3 units at 6.5% (NGSP are shown). The y-axis indicates the relative frequency of the population.