Editorial

Hui Qi Low, Andrea Rita Horvath, Tze Ping Loh*, Mario Plebani and Chun Yee Lim

Setting analytical performance specification by simulation (Milan model 1b)

https://doi.org/10.1515/cclm-2025-0121

Keywords: analytical performance specification; biological variation; clinical chemistry; quality control; external quality assurance; method evaluation

There is presently no consensus on the acceptable reclassification rate (or other clinical performance criterion), which challenges the development of model 1b analytical performance specifications (APS). Here, simulations were performed on the NHANES IV data to obtain bias APS by constraining the clinical performance criterion as an increase in reclassification rate of 1.9 % (indirect outcomes, model 1b) and flagging rate of 1.9 % (biological variation, model 2). The bias APS by model 1b were asymmetrical at upper and lower reference limits and may be lower than model 2. This may be due to the use of actual laboratory data distribution from NHANES IV (model 1b) instead of idealized assumptions without analytical variations in model 2. Nonetheless, the APS figures obtained in this exercise were meant for illustrative purposes and may not be considered definitive APS.

The Milan consensus has proposed three models for setting analytical performance specifications (APS)based on clinical outcomes (model 1), biological variation (model 2) and the state-of-the-art (model 3) [1]. Significant challenges exist in setting APS based on outcome studies (Model 1a), where the impact of the analytical performance of the test is directly observed through following clinical outcomes. Instead, the indirect outcome-based approach (Milan model 1b), which assesses the impact of analytical performance on clinical decisions related to classification of conditions, is generally considered more pragmatic and feasible. When quantitative laboratory results are interpreted against a reference limit or clinical decision limit, a change in analytical performance such as an increase in bias or imprecision can shift the patient population outside of the interpretative threshold, leading to an increase in flagging rate or reclassification, respectively.

This approach often involves statistical simulation of the impact of analytical performance on the clinical performance of a test. Simulation-based approaches in deriving APS require considerable statistical proficiency which presents a barrier to their wider adoption. Recently, Cubukcu described an electronic tool (APS Simulator) that allows to perform such analysis in a user-friendly manner [2] and that complements other existing tools [3]. Nonetheless, there is presently no consensus on the acceptable reclassification rate (or other clinical performance criterion), which challenges the development of model 1b APS [4]. In this Editorial, we compared the APS of the models 1b (indirect outcomes based on flagging or reclassifcation rates) and 2 (biological variation) when using a clinical performance criterion that statistically aligns with the biological variation model [3].

The aim of Model 1 APS is to set analytical performance criteria for tests that meet the clinical needs of patients and that make a test fit for its intended clinical use. The relationship between APS and clinical outcome is dynamic, continuous [3] and indirectly linked through the clinical performance of a test in a given clinical application or diagnostic pathway and the consequent medical decisions based on the test result [4]. Hence, it is necessary to first fix a desirable clinical performance criterion as a constraint factor and to derive a discrete numerical APS which best meets that predefined clinical performance [2, 3]. Such clinical performance criterion may include clinically acceptable reclassification (misclassification) rate, clinical sensitivity and specificity. Different clinical performance criterion may be applied depending on the purpose of the APS, the statistical approach, and the clinical utility of the measurand. Currently no generally agreed threshold for the clinical performance criterion exists for the derivation of Model 1b APS [4].

Statistical approaches that adopt acceptable reclassification rates as the clinical performance criterion, could be aligned with the biological variation model (Milan model 2)

***Corresponding author: Tze Ping Loh**, Department of Laboratory Medicine, National University Hospital, 5 Lower Kent Ridge Road, Singapore 119074, Singapore, E-mail: tploh@hotmail.com

Hui Qi Low and Chun Yee Lim, Engineering Cluster, Singapore Institute of Technology, Singapore, Singapore

Andrea Rita Horvath, Department of Chemical Pathology, Prince of Wales Hospital, Sydney, NSW, Australia

Mario Plebani, Department of Laboratory Medicine, University of Padova, Padova, Italy. https://orcid.org/0000-0002-0270-1711

[5]. Under the biological variation model, the APS for desirable bias is determined by [0.25 × √((within subject biological variation)^2 + (between-subject biological variation)^2)]. This degree of bias is associated with an increased flagging rate at the reference limits of up to 4.4 % from the theoretical 2.5 % [6]. This represents an implied acceptable reclassification rate (i.e. increased flagging rate due to the presence of an analytical bias) of 1.9 % (i.e. 4.4 % minus 2.5 %). Using the 1.9 % reclassification rate as a tolerable clinical performance criterion, the Model 1b APS was estimated in a simulation study.

The National Health and Nutrition Examination Survey (NHANES) provides data in a large community-based population in the USA. The data for 2021–2023 were retrieved (n=15,562). Participants who were more than 18 years old and did not have any clinical history or on any medications were included (n=1,571) (Supplemental Table 1). Participants who had two or more laboratory results for 14 common serum biochemistry measurands that were outside of a set of previously recommended harmonised reference intervals (Supplemental Table 2) [7, 8] were excluded, leaving 879 presumed healthy reference subjects.

Following outlier exclusion by Tukey's criteria, the 2.5th and 97.5th percentile reference limits for the 14 common biochemistry measurands of the reference subjects were calculated (Supplemental Table 2). Subsequently, analytical bias was simulated on the original laboratory values of the reference subjects through Gaussian random number generation as previously described [3]. The change in classification of the simulated results from the original value relative to the derived reference limits (i.e. reclassification rate) was recorded. This simulation was performed with 1,000 replicate runs and the averaged results are presented. The analytical bias associated with a reclassification rate of 1.9 % for the lower and upper reference limits were recorded in Table 1.

The APS for bias derived from the simulation were compared against the desirable APS for bias derived using biological variation data collated in the European Federation of Clinical Chemistry and Laboratory Medicine biological variation database (Table 1) [9]. In both instances, the impact of bias and imprecision can have asymmetrical impact on the reclassification rate at the upper and lower reference limit due to skewed distribution. Laboratory users may reconcile these differences by considering the clinical use of the measurand and select the bias APS for the reference limit that is most clinically relevant – for example, if a test is used mostly to diagnose conditions associated with increased concentration of the measurand, then the bias APS associated with upper reference limit may be preferred, and vice versa.

It is notable that the APS derived using the same clinical performance criterion of an acceptable increase of 1.9 % in flagging/reclassification rate are different for the indirect outcome-based and biological variation APS models (see Supplemental Table 3 for minimum APS for bias or acceptable/tolerable increase in flagging rate of 3.2 %). This is despite the fact that both APS models share a common underlying statistical basis – that the analytical bias is imposed on an underlying numerical distribution. An explanation for this observed difference may be related to the lack of inclusion of analytical variation in the statistical tool used in the APS model based on biological variation. This may have resulted in a simplified but idealized numerical distribution that allowed more convenient estimation of flagging rates (e.g. using Z-probability theory). On the other hand, numerical distribution can vary significantly between different measurands, population and measurement procedures in real world scenario. The application of simulated bias on actual population data is likely to produce more realistic estimation of reclassification rates. The availability of more advanced computing power and software may reduce the barrier of adopting this approach [2, 3]. The APS derived in

Table 1: Analytical performance specification (APS) for bias derived from biological variation (Milan model 2) and indirect outcome approach (Milan model 1b) using the same clinical performance criterion of an acceptable increase in flagging/reclassification rate of 1.9 %.

Measurand	Bias, %		
	Desirable APS by biological variation model	APS with 1.9 % reclassification rate at lower reference limit	APS with 1.9 % reclassification rate at upper reference limit
Sodium, mmol/L	0.2	−0.1	0.1
Potassium, mmol/L	1.6	−2.5	0.4
Chloride, mmol/L	0.4	−0.1	0.1
Bicarbonate, mmol/L	1.5	−0.1	0.2
Creatinine, µmol/L (male)	4.2	−4.5	3.8
Creatinine, µmol/L (female)	4.2	−8.4	3.6
Total calcium, mmol/L	0.8	−0.2	0.2
Phosphorus, mmol/L	3.3	−5.1	3.5
Lactate dehydrogenase, IU/L	3.1	−2.9	2.3
Alkaline phosphatase, IU/L	5.5	−5.8	5.4
Total protein, g/L	1.1	−2.2	1.8
Albumin, g/L	1.2	−1.1	0.2
Total bilirubin, µmol/L	8.0	–	2.5
Aspartate amino-transferase, U/L	5.3	−7.9	9.4
Alanine amino-transferase, U/L	9.3	−15.6	9.3

this study is meant to illustrate the difference between the biological variation and indirect outcomes (model 1b) approaches and should not be considered as definitive analytical performance requirements. Some of the APS derived in this manner may be overly narrow or easily achievable by routine measurement procedures, and other clinically more relevant flagging rates may be selected as tolerable reclassification criteria.

The use of reclassification rate as a clinical performance criterion in deriving indirect outcome-based APS provides an indication of the potential increase in the flagging rates of a test (i.e. results falling outside of a reference limit). An abnormal test result itself cannot be considered a clinical outcome as the latter commonly relates to return to health or avoidance of mortality. Nonetheless, potential diagnostic error (at least biochemically) due to increased analytical error (bias and/or imprecision) can adversely affect the process of patient recovery, by disutility of the care process and is considered a clinical outcome by some [10, 11]. For example, an erroneous tumour marker result due to significant bias can lead to unnecessary investigations, anxiety, or inappropriate treatment [12, 13].

The 'acceptable' flagging rate used in the biological variation model for deriving APS is arbitrarily determined and is the subject of ongoing debate [14]. However, any clinical performance criterion used in deriving indirect outcome-based APS is likely to be arbitrary too since the impact of analytical error on clinical outcome exists in a continuum. Ostensibly, the clinical risk of reclassifying a patient who is originally close to the edge of the reference limit is likely to be lower than for someone who is further away from the reference limit. It is important to recognize that the potential impact of analytical errors on reclassification rates should be considered in the context of the use of the biomarker in the relevant clinical pathway. Using a harmonised acceptable flagging rate aligned with the allowable bias in the biological variation model can facilitate a more consistent description of Model 1b APS and a more direct comparison of the two APS approaches.

Research ethics: Not applicable as this study does not involve human subjects.

Informed consent: Not applicable as this study does not involve human subjects.

Author contributions: 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. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: consensus statement from the 1st strategic Conference of the European Federation of Clinical Chemistry and Laboratory medicine. Clin Chem Lab Med 2015;53:833–5.
2. Çubukçu HC, Vanstapel F, Thelen M, van Schrojenstein Lantman M, Bernabeu-Andreu FA, Meško Brguljan P, et al. APS calculator: a data-driven tool for setting outcome-based analytical performance specifications for measurement uncertainty using specific clinical requirements and population data. Clin Chem Lab Med 2023;62:597–607.
3. Loh TP, Markus C, Lim CY. Impact of analytical imprecision and bias on patient classification. Am J Clin Pathol 2024;161:4–8.
4. Horvath AR, Bell KJL, Ceriotti F, Jones GR, Loh TP, Lord S, et al. Outcome-based analytical performance specifications: current status and future challenges. Clin Chem Lab Med 2024;62:1474–82.
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. Fraser CG. Biological variation: from principles to practice. Washington, DC: AACC press; 2001.
7. Tate JR, Sikaris KA, Jones GR, Yen T, Koerbin G, Ryan J, et al. Harmonising adult and paediatric reference intervals in Australia and New Zealand: an evidence-based approach for establishing a first panel of chemistry analytes. Clin Biochem Rev 2014;35:213–35.
8. Koerbin G, Sikaris K, Jones GRD, Flatman R, Tate JR. An update report on the harmonization of adult reference intervals in Australasia. Clin Chem Lab Med 2018;57:38–41.
9. Aarsand AK, Fernandez-Calle P, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, et al. The EFLM biological variation database. https://biologicalvariation.eu/ [Accessed 23 Nov 2024].
10. Porter ME. Measuring health outcomes: the outcomes hierarchy. N Engl J Med 2010;363:2477–81.
11. Plebani M. Harmonizing the post-analytical phase: focus on the laboratory report. Clin Chem Lab Med 2024;62:1053–62.
12. Loh TP, Lee LC, Sethi SK, Deepak DS. Clinical consequences of erroneous laboratory results that went unnoticed for 10 days. J Clin Pathol 2013;66:260–1.
13. Liu J, Tan CH, Badrick T, Loh TP. Moving sum of number of positive patient result as a quality control tool. Clin Chem Lab Med 2017;55:1709–14.
14. Oosterhuis WP, Coskun A, Sandberg S, Theodorsson E. Performance specifications for sodium should not be based on biological variation. Clin Chim Acta 2023;540:117221.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2025-0121).