Opinion Paper

Hikmet Can Çubukçu*, Florent Vanstapel, Marc Thelen, Marith van Schrojenstein Lantman, Francisco A. Bernabeu-Andreu, Pika Meško Brguljan, Neda Milinkovic, Solveig Linko, Mauro Panteghini and Guilaine Boursier, on behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group Accreditation, ISO/CEN Standards (WG-A/ISO)

APS calculator: a data-driven tool for setting outcome-based analytical performance specifications for measurement uncertainty using specific clinical requirements and population data

https://doi.org/10.1515/cclm-2023-0740
Received July 13, 2023; accepted October 18, 2023;
published online November 17, 2023

Abstract

Objectives: According to ISO 15189:2022, analytical performance specifications (APS) should relate to intended clinical use and impact on patient care. Therefore, we aimed to develop a web application for laboratory professionals to calculate APS based on a simulation of the impact of measurement uncertainty (MU) on the outcome using the chosen decision limits, agreement thresholds, and data of the population of interest.

Methods: We developed the "APS Calculator" allowing users to upload and select data of concern, specify decision limits and agreement thresholds, and conduct simulations to determine APS for MU. The simulation involved categorizing original measurand concentrations, generating measured (simulated) results by introducing different degrees of MU, and recategorizing measured concentrations based on clinical decision limits and acceptable clinical misclassification rates. The agreements between original and simulated result categories were assessed, and values that met or exceeded user-specified agreement thresholds that set goals for the between-category agreement were considered acceptable. The application generates contour plots of agreement rates and corresponding MU values. We tested the application using National Health and Nutrition Examination Survey data, with decision limits from relevant guidelines.

Results: We determined APS for MU of six measurands (blood total hemoglobin, plasma fasting glucose, serum total and high-density lipoprotein cholesterol, triglycerides, and total folate) to demonstrate the potential of the application to generate APS.

***Corresponding author: Hikmet Can Çubukçu**, MD, EuSpLM, General Directorate of Health Services, Rare Diseases Department, Turkish Ministry of Health, Bilkent Yerleskesi, 6001. Cadde, Universiteler Mahallesi 06800, Çankaya, Ankara, Türkiye; and Hacettepe University Institute of Informatics, Ankara, Türkiye, Phone: +903124717881, E-mail: hikmetcancubukcu@gmail.com. https://orcid.org/0000-0001-5321-9354

Florent Vanstapel, Laboratory Medicine, University Hospital Leuven, Leuven, Belgium; and Department of Public Health, Biomedical Sciences Group, Catholic University Leuven, Leuven, Belgium. https://orcid.org/0000-0001-6273-856X

Marc Thelen, SKML, Foundation for Quality Assurance in Laboratory Medicine, Nijmegen, The Netherlands; and Department of Laboratory Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands. https://orcid.org/0000-0003-1771-669X

Marith van Schrojenstein Lantman, SKML, Foundation for Quality Assurance in Laboratory Medicine, Nijmegen, The Netherlands; Department of Laboratory Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands; and Result Laboratory for Clinical Chemistry, Amphia Hospital Breda, Breda, The Netherlands. https://orcid.org/0000-0002-5454-990X

Francisco A. Bernabeu-Andreu, Servicio Bioquímica – Análisis Clínicos Hospital Universitario Puerta de Hierro, Madrid, Spain. https://orcid.org/0000-0001-7104-0200

Pika Meško Brguljan, Department of Clinical Chemistry, University Clinic for Respiratory and Allergic Deseases, Golnik, Slovenia. https://orcid.org/0000-0002-4945-6637

Neda Milinkovic, Department of Medical Biochemistry, Faculty of Pharmacy, University of Belgrade, Belgrade, Serbia. https://orcid.org/0000-0002-2641-9817

Solveig Linko, Linko Q-Solutions, Helsinki, Finland. https://orcid.org/0000-0003-3729-771X

Mauro Panteghini, Research Centre for Metrological Traceability in Laboratory Medicine (CIRME), University of Milan, Milan, Italy. https://orcid.org/0000-0002-7147-3433

Guilaine Boursier, Department of Molecular Genetics and Cytogenomics, Rare Diseases and Autoinflammatory Unit, CHU Montpellier, University of Montpellier, Montpellier, France. https://orcid.org/0000-0002-2903-3135

Conclusions: The developed data-driven web application offers a flexible tool for laboratory professionals to calculate APS for MU using their chosen decision limits and agreement thresholds, and the data of the population of interest.

Keywords: analytical performance specifications; measurement uncertainty; decision limits; outcome; data; ISO 15189

Introduction

Analytical performance specifications (APS) are used to monitor the performance of examination procedures [1], set goals for *in vitro* diagnostic medical devices [2], provide limits for external quality assessment programs [3], ensure laboratory work in conformance with acceptable quality [4], and for the management of result release [5].

Models for derivation of APS were first described according to a hierarchy in 1999 at the IFCC-IUPAC Stockholm conference [6]. Conversely, the 2014 EFLM Strategic Conference held in Milan formulated criteria for allocating measurands to different models, considering the measurand's role in the clinical decision-making process of a particular disease or clinical situation, as well as biological characteristics [7].

ISO 15189:2022 standard section 7.3.1 sub-section (b) states that "*The performance specifications for each examination method shall relate to the intended use of that examination and its impact on patient care*." [8]. However, APS that are based on biological variation and state-of-art may not fully address the intended clinical use. APSs based on the impact on clinical outcomes may be more appropriate, mainly for measurands that play a central role in diagnosis and follow-up [9]. Outcome-based studies can be classified into two categories, corresponding to direct and indirect studies [10]. Direct outcome-based studies require randomized controlled trials, which cannot be readily performed [11]. Therefore, indirect outcome-based studies have become the preferred type of study employed to define APS, most of which have been performed using different misclassification rates around clinical decision limits to assess the effect of MU on the diagnostic assignment of results [12].

Measurands are presently assigned to the Milan models by considering their impact on clinical decision-making, biological variability, and achievable analytical performance [9]. Due to the limited number of outcome-based APS [12], laboratory professionals are however inclined to use APS based on biological variation or state-of-the-art [13]. Therefore, there is a need to conduct more studies on APS based on clinical outcomes. However, it is essential to acknowledge that the APS derived from indirect studies may not be universally applicable across all laboratories, as the selected disagreement or agreement rates, clinical decision limits, and population data of concern may directly impact the calculated APS.

To address these challenges and provide laboratory professionals with a reliable and flexible tool for determining APS for measurement uncertainty (MU), our study aimed to develop a web application that enables to calculate APS using selected agreement thresholds, clinical decision limits, and data of a population of concern, thereby facilitating estimate of APS when direct or indirect outcome studies to define APS are not available. To demonstrate the data-driven web application, we tested it using National Health and Nutrition Examination Survey (NHANES) data.

Materials and methods

We developed a data-driven web application called "APS Calculator" that utilizes simulation techniques, as depicted in Figure 1, to determine APS for MU. The APS Calculator was designed to provide users a clear representation of the distribution of the original uploaded data, the APS for MU, and the visualization of the APS using contour plots.

The application allows the user to upload an Excel or csv file, select analyte data by choosing column name, specify clinical decision limits with their values, and enter agreement thresholds. After activating the "Simulate & Calculate" routine, the application draws a histogram plot to illustrate the distribution of the data (an example of histogram plot was given in Supplementary Materials). Subsequently, the tool conducts computer simulations to determine the APS for MU and visualize them through contour plots.

Simulation process

The APS calculator simulates the reanalysis of the same samples and evaluates the impact of simulated measurement uncertainty on the agreement between the original and simulated results based on clinical decision limits. This serves as a proof of concept, utilizing readily available data. However, as we will discuss later, the user is preferably encouraged to use their own data applicable to their clinical setting. The original data reflect a given distribution of patient data and serve as an equivalent distribution of true values. While these values are contaminated with inherent measurement uncertainty (MU), generally, the effect of MU is negligible compared to the distribution of patient data. Thus, during the simulation, laboratory results uploaded into the application constitute "fixed observations used as seeds for on average true values distributed over the whole clinical spectrum".

Regardless of whether laboratories (are allowed to) correct unacceptable/significant bias, there are always acceptable variations that introduce some bias, all with their own frequency. Every calibration and lot change will introduce some shift in results, which, in fact, is an acceptable bias. If judged as acceptable/negligible, such bias will not be managed and, therefore, will become part of long-term MU [14, 15]. On the other hand, if medically unacceptable bias exists, bias correction can

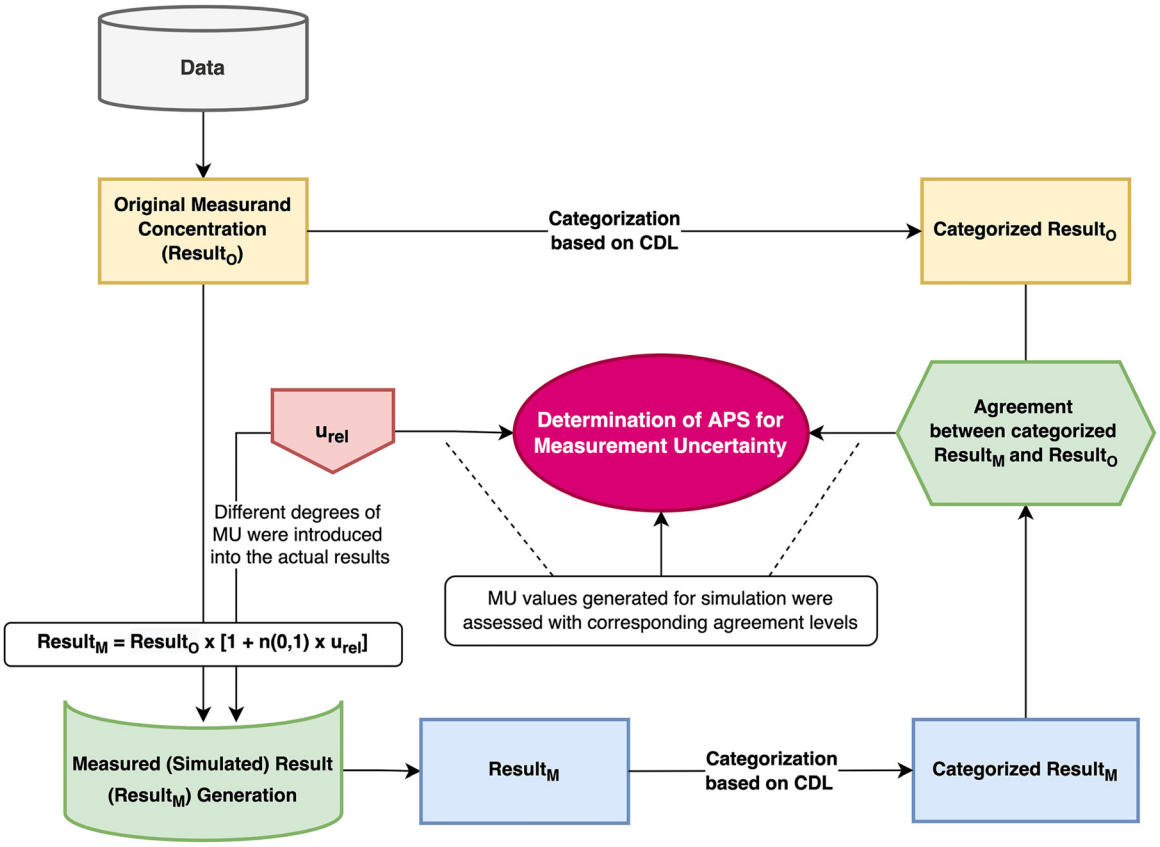

Figure 1: Summary of the process for determination of analytical performance specifications. NHANES: National Health and Nutrition Examination Survey, APS: analytical performance specifications, CDL: clinical decision limit, u_{rel}: relative standard measurement uncertainty, $Result_O$: original concentration of the measurand, $Result_M$: measured (simulated) concentration of the measurand, n(0,1): a random number generated with a normal distribution having a mean of 0 and a standard deviation of 1.

be performed. Then the uncertainty of bias correction (ubias) is calculated and incorporated in measurement uncertainty calculation [15]. Therefore, to generate simulated results influenced by MU, we used the formula proposed by Boyd and Bruns [16], modified by excluding bias from the simulation process. The modified formula is presented below:

$$Result_M = Result_O \times [1 + n(0,1) \times u_{rel}]$$

$Result_O$: Original concentration of the measurand, mostly a single estimate of the true value, $Result_M$: Measured (simulated) concentration of the measurand. After its generation, $Result_M$ is rounded up according to the number of decimals of the $Result_O$ entered by the user, n(0,1): A random number generated with a normal distribution having a mean of 0 and a standard deviation of 1, u_{rel}: Relative standard MU (in the formula, expressed as fractions).

The simulation process comprises three steps as follows:

Step 1: Categorization of $Result_O$ according to entered clinical decision limits.

Step 2: Generating measured (simulated) results by introducing u_{rel} to the original concentrations using the aforementioned formula.

Step 3: Recategorization of measured (simulated) concentration based on the clinical decision limits set in Step 1.

Note: After its generation $Result_M$ is rounded up according to the number of decimals of the $Result_O$ entered by the user.

The simulation is repeated for 331 different u_{rel} rates ranging from 0 to 33.1 % (0–0.331 in fraction unit) with intervals of 0.1 % (0.001 in fraction unit). The selection of a maximum 33 % u_{rel} limit is based on the assumption that laboratory measurements follow a Gaussian distribution. A u_{rel} exceeding 33 % would indicate a non-normal distribution of measured values for the same samples, which is impossible in replicate lab measurements where consistency is expected.

Simulation is repeated ten times for each u_{rel} rate. The generation of $Result_M$ in each repeated simulation was conducted using pseudo-random numbers (n(0,1)). The numbers (n(0,1)) within the equation were generated using the "numpy.random" module [17], which uses a pseudo-random number generator generally utilized for statistical modeling and simulation. The generator performs sampling from various probability distributions. If a seed integer is provided to the seed function, the generator performs a reproducible deterministic sampling depending on the seed value. APS calculator uses ten different seed values for the ten different simulations to ensure that each pseudo-random distribution is reproducible. The overall agreement between the $Result_O$ and $Result_M$ categories and the sublevel agreement based on the clinical decision limits are calculated for each u_{rel} rate along with sensitivity and specificity values. The final metrics values are obtained by calculating the mean values of the metrics determined from each u_{rel} rate. As a result, the APS calculator generates reproducible results for the same experiment settings.

Classification performance metrics

Following the generation of $Result_M$, the application proceeds to compute classification performance metrics for both $Result_O$ and $Result_M$. The metrics of overall agreement, sublevel agreement, sublevel sensitivity, and sublevel specificity were determined using the following formulas:

Overall agreement : TPs / Total Number of Results
Sublevel agreement : (TP + TN) / (TP + TN + FP + FN)
Sublevel sensitivity : TP / (TP + FN)
Sublevel specificity : TN / (TN + FP)

TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Determination of analytical performance specifications

u_{rel} values that correspond to an agreement level greater than or equal to the value entered by the user in the minimum agreement threshold box are considered acceptable. Among the acceptable u_{rel} values, the APS Calculator determines the minimum value of u_{rel} that is greater than or equal to the minimum agreement level entered by the user, in order to obtain the 'minimum' APS. To determine the desirable and optimal APS of u_{rel}, the values entered in the desirable and optimal agreement threshold boxes are used, in the same manner as for determining the minimum APS.

Contour plots

The web application produces a series of interactive contour plots that visually represent agreement rates as percentages on the x-axis and their corresponding u_{rel} values on the y-axis (see Figure 2 as an example). Vertical dashed lines are added to indicate the minimum, desirable, and optimal APS for u_{rel} (%), corresponding to different agreement levels. These interactive contour plots provide a clear illustration of the APS for u_{rel} (%) for both overall agreement and sub-level agreement. Additionally, the sublevel contour plots present sensitivity and specificity values, effectively showcasing the inherent trade-offs.

Test data

We used publicly available NHANES 2017–2020 pre-pandemic data, which constitutes a representative sample of the United States population [18], to test our web application. Laboratory procedures manuals of NHANES data were given as URL links in the Supplementary Materials. National Center for Health Statistics approved the survey (protocol #2018-01) [19], and informed consent for interview, specimen storage, and continuing studies were taken [20]. First, we combined some items of the questionnaire (medical history), demographics (age, sex, ethnic group), physical examination (body mass index), and laboratory data of six measurands (blood total hemoglobin (Hb), plasma glucose, serum total and high-density lipoprotein (HDL) cholesterol, triglycerides, and total folate). We then excluded subjects having missing data from the included items of the questionnaire, examination and demographic data. In addition, we excluded individuals under the age of 18 and over 65 years. The final enrollment comprised data of 5,708 subjects with the same number of results for included laboratory measurands except for plasma glucose, blood total Hb, and serum total folate. For glucose, we included only the results of the subjects who fasted for >8 h to categorize results according to fasting plasma glucose (FPG) decision limits for diabetes mellitus diagnosis. As pregnancy status was not available for all women, we included only male Hb data in the simulation process related to this measurand. Indeed, decision limits for blood Hb concentrations are dependent on the pregnancy status [21]. Finally, we could only include 3,749 total folate results due to the missing values in a minority of subjects.

Selection of decision limits

Decision limits of FPG were obtained from American Diabetes Association's recommendations [22]. National Cholesterol Education Program Expert Panel's proposed classifications based on coronary heart disease risk were used as decision limits of serum total cholesterol, HDL, and triglycerides [23]. The World Health Organization's recommendations for serum folate (based on macrocytic anemia as an indicator) [24] and blood Hb levels (based on anemia) [21] were utilized.

Demonstration

To demonstrate the utility of the APS Calculator, we used laboratory results from NHANES for the six selected measurands. The data was uploaded to the web application, and the clinical decision limits mentioned earlier were entered into their respective input widgets. The minimum, desirable, and optimal agreement thresholds were arbitrarily set at 90 %, 95 %, and 99 %, respectively. Subsequently, we introduced u_{rel} into the actual concentration of the measurand using the APS Calculator, simulating "measured" values. We then classified the actual results of selected measurands, along with their simulated values, based on their decision limits. Finally, we determined the APS for u_{rel} (%) using the approach described in the previous sections.

Python script, packages and deployment

Data pre-processing, simulation, and APS determination processes were carried out using Python version 3.9 programming language [25] and Microsoft Visual Studio Code version 1.77.3. Data processing steps were performed using Pandas version 1.3.5 [26]. Some of the mathematical functions utilized in the MU simulation were conducted using Scikit-learn 1.2.2 metrics package [27]. Histogram plots and contour plots were illustrated using Plotly version 5.15.0 [28]. Pseudo-random number generation was performed using NumPy version 1.21.2 [17]. Streamlit version 1.21.0 was used to develop web application and its deployment [29].

Results

Table 1 depicts relevant demographic characteristics of the NHANES population and the main laboratory test data. Table 2 shows the minimum, desirable, and optimal APS values for u_{rel} (%) of the six measurands as obtained by using NHANES data. In some cases, APS values belonging to different agreement subgroups of the same measurand

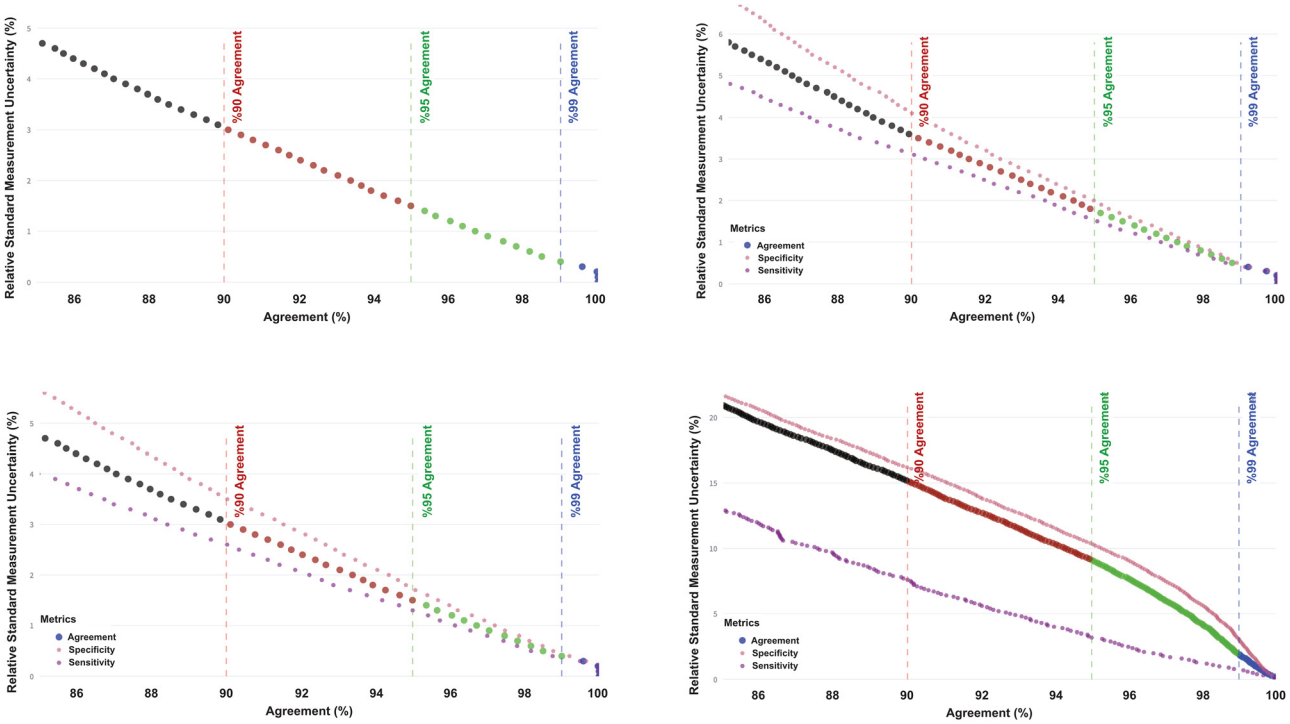

Figure 2: Contour plots defining analytical performance specifications (APS) for fasting plasma glucose relative standard measurement uncertainty (u_{rel}) based on national health and nutrition examination survey data as an example. (A) Contour plot of APS for u_{rel} (%) based on overall agreement. (B) Contour plot of APS for u_{rel} (%) based on the agreement of subgroup "<5.6 mmol/L (<100 mg/dL)" (desirable concentrations). (C) Contour plot of APS for u_{rel} (%) based on the agreement of subgroup "5.6–6.9 mmol/L (100–125 mg/dL)" (impaired fasting glucose). (D) Contour plot of APS for u_{rel} (%) based on the agreement of subgroup "≥7 mmol/L (≥126 mg/dL)" (diabetes mellitus). Dashed vertical lines represent the selected agreement thresholds. 90 % agreement represents the minimum APS, the 95 % agreement the desirable APS, and the 99 % agreement the optimum APS, respectively. Black, red, green, and blue colored agreement dots correspond to unacceptable, minimum, desirable, and optimal u_{rel} value ranges.

displayed significant differences and it could be difficult to apply different APS to the same measurand. As a starting practical point, APS values based on overall agreement may provide more balanced goals encompassing all subgroups. As an example, Figure 2 displays the APS contour plots of FPG. To fulfill APS derived using our model, u_{rel} of FPG should not exceed 3.0 and 1.4 % at minimum and desirable quality level, respectively.

The APS Calculator's Python scripts have been made publicly available through our GitHub repository at https://github.com/hikmetc/APSCalculator. Furthermore, the APS Calculator can be easily accessed via any web browser using the following URL: https://hikmetc-apscalculator.streamlit.app.

Discussion

In this study, we have developed a data-driven web application for helping laboratory professionals in calculating APS for MU using their own selected agreement thresholds, clinical decision limits, and data of a population of concern. To demonstrate the use of our application, we utilized NHANES data as an example. Our proposal should be considered as an indirect outcome-based model as described in the 2014 EFLM Milan Conference consensus [10].

Various approaches have been proposed to derive indirect outcome-based APS, which can be summarized under obtaining "true" values, generating measured values, and assessing the effect of measurement differences on outcomes [12]. One of the most used approaches for generating measured results is the formula proposed by Boyd and Bruns [12, 16]. Originally, this formula considered imprecision and bias as two distinct error sources. However, according to the implementation of metrological traceability theory, one should act to eliminate a "medically significant" measurement bias through appropriate correction of calibration and then estimate MU at the commercial calibrator and clinical sample levels [30, 31]. Furthermore, when the imprecision component is calculated over a long period, including

Table 1: Main demographic characteristics and laboratory features of national health and nutrition examination survey (NHANES) population (n=5,708).

Age (years) at screening, median (25–75th)		**45 (32–56)**
Sex, n (%)		
Female		3,041 (53.3 %)
Male		2,667 (46.7 %)
Ethnic group, n (%)		
Mexican American		777 (13.6 %)
Other Hispanic		642 (11.2 %)
Non-Hispanic White		1,734 (30.4 %)
Non-Hispanic Black		1,513 (26.5 %)
Non-Hispanic Asian		749 (13.1 %)
Others		293 (5.1 %)
Body mass index, kg/m^2, median (25–75th)		29 (24.9–34.2)
Medical history, n (%)		
Angina pectoris		86 (1.5 %)
Arthritis		1,298 (22.7 %)
Asthma		960 (16.8 %)
Malignancy		315 (5.5 %)
Congestive heart failure		110 (1.9 %)
COPD, emphysema, chronic bronchitis		402 (7 %)
Coronary heart disease		118 (2.1 %)
Diabetes mellitus		654 (11.5 %)
Gallstones		495 (8.7 %)
Heart attack		145 (2.5 %)
Hepatitis B		65 (1.1 %)
Hepatitis C		98 (1.7 %)
Hypertension		1,741 (30.5 %)
Chronic kidney failure		143 (2.5 %)
Stroke		168 (2.9 %)
Dietary supplements usage (in the past month)		2,917 (51.1 %)
Taken prescribed drugs (in the past month)		2,953 (51.7 %)
Taking prescribed drugs for hypercholesterolemia		789 (66.9 %)
Smoking history (at least 100 cigarettes in life)		2,260 (39.6 %)
Fasting duration, h, median (25–75th)		10 (5–12)
Measurand, n, median (25–75th)		
Fasting plasma glucose, mmol/L – mg/dL	2,758	5.7 (5.3–6.2) – 102 (95–112)
S-Total cholesterol, mmol/L – mg/dL	5,708	4.76 (4.14–5.48) – 184 (160–212)
S-HDL-cholesterol, mmol/L – mg/dL	5,708	1.29 (1.09–1.58) – 50 (42–61)
S-Triglycerides, mmol/L – mg/dL	5,708	1.24 (0.87–1.85) – 110 (77–164)
B-Hemoglobin, g/L	2,667	150 (143–157)
S-Total folate, µg/L	3,749	13.7 (9.5–20.2)

COPD, chronic obstructive pulmonary disease.

different sources of systematic variation, such as calibrations, reagent lot changes, and instrument maintenance, it will include all bias associated with such variations [14]. Therefore, we modified the formula offered by Boyd and Bruns by keeping only u_{rel} as a source of measurement variability. Furthermore, when there is a bias relative to the method used for determining clinical decision limits, bias correction can be performed (if permitted). In such circumstances, the laboratory may calculate the uncertainty of bias correction (ubias) and combine it with other measurement uncertainty components to achieve relative standard uncertainty of a measured value. The user can input this relative standard uncertainty into the APS calculator, which subsequently conducts further simulations per the provided parameters.

It is important to note that the formula employed in this study assumes the absence of pre-analytical sources of error.

Table 2: Analytical performance specifications (APS) for relative standard measurement uncertainty (u_{rel} (%)) derived using the APS Calculator as described in this paper.

Measurand	Agreement groups	APS for u_{rel}, %		
		Minimum (90 % agreement)	Desirable (95 % agreement)	Optimal (99 % agreement)
S-Triglycerides	Overall agreement	12.4	6.1	1.2
	Subgroup agreement			
	Desirable: <1.7 mmol/L (<150 mg/dL)	22.0	10.5	2.0
	Borderline high: 1.7–2.2 mmol/L (150–199 mg/dL)	13.0	6.3	1.2
	High: 2.3–5.6 mmol/L (200–499 mg/dL)	31.5	15.0	2.9
	Very high: ≥5.6 mmol/L (≥500 mg/dL)	NA[a]	NA[a]	NA[a]
S-Total cholesterol	Overall agreement	4.8	2.3	0.4
	Subgroup agreement			
	Desirable: <5.17 mmol/L (<200 mg/dL)	7.2	3.5	0.6
	Borderline high: 5.17–6.18 mmol/L (200–239 mg/dL)	4.8	2.3	0.4
	High: ≥6.21 mmol/L (≥240 mg/dL)	14.9	7.3	1.4
S-HDL-cholesterol	Overall agreement	5.6	2.9	0.8
	Subgroup agreement			
	Low: <1.03 mmol/L (<40 mg/dL)	11.8	5.9	1.5
	Desirable: 1.03–1.54 mmol/L (40–59 mg/dL)	5.6	2.9	0.8
	High: ≥1.55 mmol/L (≥60 mg/dL)	11.3	5.4	1.1
P-Fasting glucose	Overall agreement	3.0	1.4	0.3
	Subgroup agreement			
	Desirable: <5.6 mmol/L (<100 mg/dL)	3.5	1.7	0.4
	IFG: 5.6–6.9 mmol/L (100–125 mg/dL)	3.0	1.4	0.3
	Diabetes mellitus: ≥7.0 mmol/L (≥126 mg/dL)	15.1	9.1	1.9
S-Total folate	Overall agreement	14.8	7.6	1.5
	Subgroup agreement			
	Deficiency: <3.0 µg/L	NA[a]	NA[a]	26.9
	Possible deficiency: 3.0–5.9 µg/L	NA[a]	20.9	5.7
	Unlikely deficiency: 6.0–20 µg/L	15.2	7.7	1.5
	Elevated: >20 µg/L	23.0	10.7	2.0
B-Hemoglobin	Overall agreement	8.5	5.6	1.6
	Subgroup agreement			
	Severe anemia: <80 g/L	NA[a]	27.9	19.0
	Moderate anemia: 80–109 g/L	21.2	15.1	8.5
	Mild anemia: 110–129 g/L	8.8	5.6	1.6
	Non-anemia: ≥130 g/L	8.9	5.8	1.7

[a]The low number of results in this subgroup resulted in an inaccurate APS estimate. NA, not available; IFG, impaired fasting glucose; u_{rel}, relative standard MU (expressed in percentage units).

However, to the best of our knowledge, previous research has also not accounted for pre-analytical variation. On the other hand, APS calculator simulates "re-analysis of the same samples," which focuses only on the analytical phase; therefore, intraindividual biological variation was excluded from the simulation process due to the absence of re-sampling.

The effects of measurement differences on clinical outcome can be assessed within the different scopes of clinical accuracy, treatment decision, and costs [12]. Most of the indirect outcome studies in the current literature evaluated the clinical accuracy using clinically acceptable misclassification rates as an identifier of APS [9, 12, 13]. However, different misclassification rates were employed in these studies. For instance, Boyd and Bruns used an error rate of 5 % in insulin dosing for glucometers [16]. Similarly, Von Eyben et al. utilized 5 % false result rates for lactate dehydrogenase isoenzyme 1 [32]. Stöckl et al. considered a misclassification of 20 % as an acceptable rate to determine APS of vitamin D [33]. Finally, Ferraro et al. used

misclassified results of 0.5–2.5 % to identify APS of serum folate [34]. Overall, there has not been any consensus on the amount of misclassification rate for determining APS in indirect outcome studies, even because the acceptable misclassification is of course dependent on the type of disease to be diagnosed and the harm inflicted by misclassification. Thus, to provide a flexible tool, we left agreement thresholds (i.e., the reverse form of the misclassification rate) as an input to be determined and entered by users.

In the literature, the assessment of the effect of MU on outcomes has been conducted using contour plots that incorporate both sensitivity and specificity values. However, it should be noted that these plots, as demonstrated in a hypothetical example by Smith et al. [12], primarily focus on discrimination between healthy and patient populations, which corresponds to a 2×2 confusion matrix where the calculation of sensitivity and specificity values is straightforward. In our study, we opted to include only overall agreement, sublevel agreement, sublevel sensitivity, and sublevel specificity in the assessment of classification performance due to the intrinsic characteristics of confusion matrix statistics. When a single clinical decision limit is employed, the confusion matrix represents the categories of the uploaded data in a 2×2 formation. In this scenario, the overall sensitivity, overall specificity, and overall agreement are equivalent to each other, while the sensitivity of one sublevel is equivalent to the specificity of another sublevel. However, when there are two or more clinical decision limits, the confusion matrix typically represents a 3×3 or larger formation. In such cases, stable true negatives are absent, resulting in an overall specificity value of zero. Nonetheless, the overall agreement and overall sensitivity remain equal. As a result, to address this confusion, we have excluded the overall sensitivity and overall specificity from the contour plots depicting the overall APS for standard MU and from the overall APS determination process. Instead, the sublevel contour plots illustrate the sublevel sensitivity and sublevel specificity to demonstrate the trade-offs.

The significance of the variations in APS reported in indirect outcome studies can be attributed to the utilization of different data, various decision limits, several misclassification rates, and diverse calculation methods. Loh et al.'s literature review concerning APS in the context of HbA_{1c} revealed substantial discrepancies not only between results obtained using different models but also within studies using only indirect outcome approaches, with APS ranging from 2 to 25 % [35]. The reported variability in APS underscores the need for a standardized and flexible approach to set APS for standard MU. In this regard, our data-driven web application can be a valuable tool for laboratory professionals. The APS Calculator provides the flexibility to use various data sources relevant to the specific clinical setting of interest. This application also enables users to customize decision limits and agreement rates based on their particular requirements, resulting in APS that are tailored and fit for purpose. In addition to its customization features, the APS Calculator can also foster transparency and reproducibility in APS determination. The application allows for the documentation and sharing of output data and parameters among researchers and stakeholders, thereby enhancing the ability to replicate results and facilitate cross-validation. Overall, the utilization of this application can improve the understanding of reported APS values, ultimately enhancing the quality of laboratory testing and patient care.

NHANES data, a comprehensive and representative sample of the United States population, consists of both healthy and unhealthy individuals across various age groups and demographic characteristics. By utilizing this publicly available dataset and internationally recommended decision limits, we were able to demonstrate the potential of our application to generate APS for various measurands along with visualisation features, as shown in Figure 2 for FPG. However, the APS identified in the current study cannot be extrapolated to all clinical scenarios. For example, if the goal is to adjust insulin dosing, different clinical decision limits may be employed to determine the APS for glucose [16].

Authors previously stressed that the Milan models use different principles and do not constitute a hierarchy [7, 9, 13, 36]. Therefore, some models are better suited for certain measurands than for others, and the attention should primarily direct toward the measurand and its biological and clinical characteristics. Particularly, the outcome-based model applies for measurands that have a central and well-defined role in the decision-making of a specific disease or clinical situation, and test results should be interpreted through established decision limits. This applies to all six selected measurands in our study. For some of them (FPG, total Hb, total cholesterol, and total folate), some information about APS using the indirect outcome model was already available [4, 34, 37–40]. For the remaining two (triglycerides and HDL), a recent paper was unable to retrieve peer-reviewed literature for outcome studies dealing with their clinical use and evaluating the impact of analytical variability on clinical outcomes [13]. Therefore, the authors temporarily allocated those measurands to the biological variation model to derive their APS. Table 3 compares previously available APS with those defined in the present study. It is noteworthy that FPG, folate, triglycerides, and HDL show very similar APS, even if the two mentioned lipid parameters have APS derived using different Milan models, i.e., indirect outcome-based in this study and biological variation-based in a previous one [13]. The similarity of APS

Table 3: Analytical performance specifications (APS) for relative standard measurement uncertainty (u_{rel}) on clinical samples for the measurands allocated to the Milan model 1 (outcome-based APS) considered in the present study. Comparison of the obtained results (under overall agreement group) with previously published data.

Measurand	Previously published APS for u_{rel}, %		APS for u_{rel} from present study, %	
	Desirable	Minimum	Desirable	Minimum
P-Fasting glucose	2.0[a]	3.0[a]	1.4	3.0
B-Hemoglobin	2.8[a]	4.2[a]	5.6	8.5
S-Total cholesterol	3.0[a]	7.0[a]	2.3	4.8
S-HDL cholesterol	2.8[a]	4.3[a]	2.9	5.6
S-Triglycerides	9.9[a]	14.9[a]	6.1	12.4
S-Total folate	8.0[b]	12.0[b]	7.6	14.8

Data from [a]ref. [4] and [b]ref. [34], respectively.

derived from both models for the two lipids seems to confirm the equivalence of the two models advocated by the expert group of the EFLM when measurands have a steady state in the blood of (normolipemic) individuals. For other two measurands, derived APS appear to be different, and some explanations may be put forward. For total Hb, only one previous paper provided information about APS by accepting a false positive rate of 5 % in classifying patients. However, this was done with some methodological limitations, so APS results cannot be definitively compared [39]. For total cholesterol, Petersen and Klee evaluated the influence of analytical variability of measurements on the number of low-risk individuals misclassified as false positive, i.e., individuals at high risk [40]. The reported APS were obtained by accepting a misclassification rate, in terms of false positive results, of approximately 5.0 % (desirable) and 7.5 % (minimum) at the 6.21 mmol/L (240 mg/dL) cut-off. Checking their simulated data in Supplementary Table 2A, the rates of subject misclassification did, however, not change markedly if our APS from the overall agreement group were considered: an APS of 2.3 % (desirable) corresponded to about 5 % false positives, and an APS of 4.8 % (minimum) to approximately 6 % of wrongly classified individuals. Applying APS specifically derived in our study for the similar agreement subgroup (decision limit ≥6.21 mmol/L [240 mg/dL]) appears, however, to amplify the noticed difference, asking for less demanding analytical goals (Table 2).

The APS calculator is designed to estimate APS by considering the entered minimum, desirable, and optimal agreement thresholds and the data provided by the user. However, it is important to consider that when the data related to a specific subgroup contains a low number of results, the estimated APS for standard MU can be inaccuratety high, exceeding the goal of 33 %, which in Table 2 were labeled as “not available”. Even if the APS calculator ensured the reproducibility of the same experiment settings with the same data, a larger laboratory data set of the same experiment settings might result in more reliable estimates of APS for standard MU and aid in overcoming inaccurately high APS for MU due the low sample size problem. Nevertheless, taking into account that overall APS for standard MU encompass all sublevels based on clinical decision limits, their related values may provide a more comprehensive set of goals for each measurand.

Conclusions

The development of a data-driven web application presented in this study may offer a valuable tool to laboratory professionals to calculate APS for standard MU using their own agreement thresholds, clinical decision limits, and, most importantly, data of the population of interest to address actual clinical questions in specific scenarios. In addition, the application can foster transparency and reproducibility in APS estimates and increase the use of APS obtained using EFLM-recommended models. Overall, the application provides a standardized and flexible approach to set APS for MU, enabling users to customize decision limits and agreement rates based on their particular requirements, resulting in APS that are tailored and fit for purpose, as specified by ISO 15189:2022 [8]. By using this tool, laboratories can define their MU as analytically acceptable on the basis of the clinical decision uncertainty personalized by using local data which do right to their own clinical setting.

Research ethics: We used publicly available NHANES 2017–2020 pre-pandemic data, which constitutes a representative sample of the United States population [17], to test our web application. National Center for Health Statistics approved the survey (protocol #2018-01) [18].

Informed consent: We used publicly available NHANES 2017–2020 pre-pandemic data. Informed consent for interview, specimen storage, and continuing studies were taken [19] for NHANES 2017–2020 pre-pandemic data.

Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest.

Research funding: None declared.

Data availability: Not applicable.

References

1. Braga F, Pasqualetti S, Aloisio E, Panteghini M. The internal quality control in the traceability era. Clin Chem Lab Med 2020;59:291–300.
2. Braga F, Panteghini M. Defining permissible limits for the combined uncertainty budget in the implementation of metrological traceability. Clin Biochem 2018;57:7–11.
3. Braga F, Pasqualetti S, Panteghini M. The role of external quality assessment in the verification of in vitro medical diagnostics in the traceability era. Clin Biochem 2018;57:23–8.
4. Panteghini M. Redesigning the surveillance of in vitro diagnostic medical devices and of medical laboratory performance by quality control in the traceability era. Clin Chem Lab Med 2023;61:759–68.
5. Çubukçu HC, Vanstapel F, Thelen M, Bernabeu-Andreu FA, van Schrojenstein Lantman M, Brugnoni D, et al. Improving the laboratory result release process in the light of ISO 15189:2012 standard. Clin Chim Acta 2021;522:167–73.
6. Kenny D, Fraser CG, Petersen PH, Kallner A. Consensus agreement. Scand J Clin Lab Invest 1999;59:585.
7. Ceriotti F, Fernandez-Calle P, Klee GG, Nordin G, Sandberg S, Streichert T, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM strategic conference. Clin Chem Lab Med 2017;55:189–94.
8. International Organization for Standardization. ISO 15189:2022 Medical laboratories – Requirements for quality and competence. Geneva, Switzerland. https://www.iso.org/standard/76677.html [Accessed 1 Sep 2023].
9. Braga F, Panteghini M. Performance specifications for measurement uncertainty of common biochemical measurands according to Milan models. Clin Chem Lab Med 2021;59:1362–8.
10. 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.
11. 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.
12. 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.
13. Braga F, Pasqualetti S, Borrillo F, Capoferri A, Chibireva M, Rovegno L, et al. Definition and application of performance specifications for measurement uncertainty of 23 common laboratory tests: linking theory to daily practice. Clin Chem Lab Med 2023;61:213–23.
14. van Schrojenstein Lantman M, Çubukçu HC, Boursier G, Panteghini M, Bernabeu-Andreu FA, Milinkovic N, et al. An approach for determining allowable between reagent lot variation. Clin Chem Lab Med 2022;60: 681–8.
15. International Organization for Standardization. ISO/TS 20914:2019 Medical laboratories – Practical guidance for the estimation of measurement uncertainty. Geneva, Switzerland. https://www.iso.org/standard/69445.html [Accessed 1 Sep 2023].
16. Boyd JC, Bruns DE. Quality specifications for glucose meters: assessment by simulation modeling of errors in insulin dose. Clin Chem 2001;47:209–14.
17. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020; 585:357–62.
18. Centers for Disease Control and Prevention. NHANES 2017-March 2020 pre-pandemic data. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020 [Accessed 1 Oct 2022].
19. NCHS Ethics Review Board (ERB) Approval: 2022. Available from: https://www.cdc.gov/nchs/nhanes/irba98.htm.
20. Centers for Disease Control and Prevention. NHANES 2017-March 2020 pre-pandemic brochures and consent documents. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/documents.aspx?Cycle=2017-2020 [Accessed 1 Aug 2022].
21. World Health Organization. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. Geneva: World Health Organization; 2011.
22. American Diabetes Association. Standards of medical care in diabetes-2022 abridged for primary care providers. Clin Diabetes 2022;40:10–38.
23. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). JAMA 2001;285:2486–97.
24. World Health Organization. Serum and red blood cell folate concentrations for assessing folate status in populations. Geneva: World Health Organization; 2015.
25. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.
26. The pandas development team. pandas-dev/pandas: Pandas. Meyrin, Switzerland: Zenodo; 2020.
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12: 2825–30.
28. Plotly Technologies Inc. Collaborative data science publisher. Montreal, QC: Plotly Technologies Inc.; 2015. https://plot.ly [Accesed 21 Jun 2023].
29. Streamlit. A faster way to build and share data apps. https://streamlit.io/ [Accesed 12 Apr 2023].
30. Panteghini M, Braga F. Implementation of metrological traceability in laboratory medicine: where we are and what is missing. Clin Chem Lab Med 2020;58:1200–4.
31. Braga F, Panteghini M. The utility of measurement uncertainty in medical laboratories. Clin Chem Lab Med 2020;58:1407–13.
32. von Eyben FE, Petersen PH, Blaabjerg O, Madsen EL. Analytical quality specifications for serum lactate dehydrogenase isoenzyme 1 based on clinical goals. Clin Chem Lab Med 1999;37:553–61.
33. Stöckl D, Sluss PM, Thienpont LM. Specifications for trueness and precision of a reference measurement system for serum/plasma 25-hydroxyvitamin D analysis. Clin Chim Acta 2009;408:8–13.
34. 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.
35. Loh TP, Smith AF, Bell KJL, Lord SJ, Ceriotti F, Jones G, et al. Setting analytical performance specifications using HbA1c as a model measurand. Clin Chim Acta 2021;523:407–14.
36. Panteghini M, Ceriotti F, Jones G, Oosterhuis W, Plebani M, Sandberg S. Strategies to define performance specifications in laboratory medicine: 3 years on from the Milan strategic conference. Clin Chem Lab Med 2017;55:1849–56.

37. Petersen PH, Brandslund I, Jørgensen L, Stahl M, de Fine Olivarius N, Borch-Johnsen K. Evaluation of systematic and random factors in measurements of fasting plasma glucose as the basis for analytical quality specifications in the diagnosis of diabetes. 3. Impact of the new WHO and ADA recommendations on diagnosis of diabetes mellitus. Scand J Clin Lab Invest 2001;61:191–204.
38. 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.
39. Thue G, Sandberg S, Fugelli P. Clinical assessment of haemoglobin values by general practitioners related to analytical and biological variation. Scand J Clin Lab Invest 1991;51:453–9.
40. 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.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2023-0740).