Anna Carobene*, Andrea Campagner, Christian Uccheddu, Giuseppe Banfi, Matteo Vidali and Federico Cabitza

The multicenter European Biological Variation Study (EuBIVAS): a new glance provided by the Principal Component Analysis (PCA), a machine learning unsupervised algorithms, based on the basic metabolic panel linked measurands

https://doi.org/10.1515/cclm-2021-0599
Received May 19, 2021; accepted July 20, 2021;
published online August 2, 2021

Abstract

Objectives: The European Biological Variation Study (EuBIVAS), which includes 91 healthy volunteers from five European countries, estimated high-quality biological variation (BV) data for several measurands. Previous EuBIVAS papers reported no significant differences among laboratories/population; however, they were focused on specific set of measurands, without a comprehensive general look. The aim of this paper is to evaluate the homogeneity of EuBIVAS data considering multivariate information applying the Principal Component Analysis (PCA), a machine learning unsupervised algorithm.

Methods: The EuBIVAS data for 13 basic metabolic panel linked measurands (glucose, albumin, total protein, electrolytes, urea, total bilirubin, creatinine, phosphatase alkaline, aminotransferases), age, sex, menopause, body mass index (BMI), country, alcohol, smoking habits, and physical activity, have been used to generate three databases developed using the traditional univariate and the multivariate Elliptic Envelope approaches to detect outliers, and different missing-value imputations. Two matrix of data for each database, reporting both mean values, and "within-person BV" (CV_P) values for any measurand/subject, were analyzed using PCA.

Results: A clear clustering between males and females mean values has been identified, where the menopausal females are closer to the males. Data interpretations for the three databases are similar. No significant differences for both mean and CV_Ps values, for countries, alcohol, smoking habits, BMI and physical activity, have been found.

Conclusions: The absence of meaningful differences among countries confirms the EuBIVAS sample homogeneity and that the obtained data are widely applicable to deliver APS. Our data suggest that the use of PCA and the multivariate approach may be used to detect outliers, although further studies are required.

Keywords: biological variation; EuBIVAS; machine learning; preanalytical phase.

***Corresponding author: Anna Carobene**, Laboratory Medicine, IRCCS San Raffaele Scientific Institute, Via Olgettina 60, 20132 Milan, Italy, Phone: +390226432850, E-mail: carobene.anna@hsr.it
Andrea Campagner, Christian Uccheddu and Federico Cabitza, University of Milano-Bicocca, Milan, Italy
Giuseppe Banfi, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy; and Università Vita e Salute San Raffaele, Milan, Italy
Matteo Vidali, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

Introduction

Estimating Biological Variation (BV) has many applications: it is used for the setting of the analytical performance specification (APS) for both internal and external quality control [1–3]; for the assessment of significance of change in serial measurements in a subject (reference change value; RCV); and for determining the number of samples needed to estimate the homeostatic setting point [4, 5]. Moreover, BV estimates are used to assess the utility of conventional population-based reference intervals through the use of the individuality index, and to derive personalized reference intervals, which can improve diagnostic accuracy and treatment appropriateness on a more patient-centered level [6].

The importance of BV estimates, the within-subject BV (CV_I) and the between-subject BV (CV_G), in the last years has been put forward through several activities carried out by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) [7].

As an outcome of the first Strategic Conference of the EFLM held in Milan in November 2014, the EFLM Task

Force and Finish Group on BV database (TFG-BVD), was established with the objective to appraise the quality of BV data that are publicly available [8]. Furthermore, responding to the need for reliable BV estimates delivered from big high-quality studies [9, 10], the EFLM WG-BV has designed and implemented the European Biological Variation Study (EuBIVAS) [11, 12].

The design of a multicenter study, aimed at obtaining reliable BV data, requires the involvement of many individuals and a large number of blood collections to provide adequate statistical power, as demonstrated by Røraas et al. [13]. The EuBIVAS is a high-power study, which included 91 apparently healthy volunteers from 5 European countries, from which high-quality BV data, based on the newest generation of analytical methods, have been derived for a set of measurands [12, 14].

However, the multicenter project thus conceived posed the risk of introducing factors impacting on the reliability and hence on the applicability of the obtained BV data. Namely, the two most important aspects were: considering the enrolled population as homogeneous, even if coming from different countries; and introducing a significant pre-analytical variability, having performed the collections in several centres.

To minimize these latter aspects, a very detailed protocol was prepared, agreed upon, and rigorously followed by each involved laboratory in all steps [14], by taking into account all the requirements of the EFLM checklist for BV studies [15, 16].

Previously published EuBIVAS papers [17–20] reported no significant differences in pre-analytical variables or treatment between the different laboratories. However, these previous works were focused on the analysis of specific set of measurands, without a comprehensive general look that could take into consideration multivariate information coming from a panel of several measurands, possibly together with some characteristics of the population (i.e. lifestyle).

Machine learning (ML) is a methodological approach developed within the field of artificial intelligence (AI) aimed at developing algorithms and models by which to make classifications and predictions on the basis of large amounts of data taken as examples to "learn" relevant patterns and schemas. The essence and main novelty of ML methods is to infer new sets of rules or new algorithms from data to perform complex cognitive tasks (like diagnosis or prognosis), without step-by-step instructions explicitly given in advance by the programmer [21]. ML is an umbrella term to denote a large variety of statistical methods, which include supervised and unsupervised approaches. In the former case, labelled data (data containing both the input and the output variables) are given to the learning algorithm to automate classification (predicting a categorical variable) or regression (predicting a continuous variable) tasks. By contrast, in the unsupervised case, only input data are given, and the goal is to try to modelling the underlying structure or distribution of the data, and to find unknown patterns. Among unsupervised algorithms, Principal Component Analysis (PCA) is one of the most widely used methods. In short, PCA is a techniques for dimensionality reduction that is an approach by which multidimensional data can be reduced into fewer, new dimensions that embed some information from the original ones. In particular, PCA describes the standardized multidimensional data in terms of components that are linear combinations of the original variables preserving the variance of the original data. For these reasons, PCA can also be applied to easily and meaningfully visualize high dimensional data in case the first 2 o 3 principal components are chosen, that is those that explain a maximal amount of variance.

In any case, whether supervised or unsupervised, accurate predictions by ML algorithms rely on data completeness and quality [21], according to the principle commonly described as "garbage in, garbage out".

In particular, the presence of outliers, or extreme values, in a dataset may result in poor predictive modelling performance. However, identifying and removing outliers can be a very challenging and time consuming task, even with a good understanding of the structure of the data. Several strategies to detect outliers are available, including univariate and multivariate methods. With univariate methods, each variable is considered independently, and a value is labelled as an outlier when it is too far (according to some criteria) from a central tendency indicator. However, these methods usually fail in detecting observations that deviate from global behavior or from the pattern of the majority of data (with pattern here we indicate the relationship between multiple variables of the dataset) [22].

In this work, by applying PCA, we evaluated the homogeneity of EuBIVAS data through a novel approach, with the goal of detecting previously unknown patterns or relationships between variables. Moreover, traditional univariate approach to detect outliers was compared to the multivariate approach. With this aim, the EuBIVAS measurands included in the basic metabolic panel (BMP), glucose (Glu), calcium (Ca) [17], albumin (Alb), total protein (TP) [18], sodium (Na), potassium (K), chloride (Cl), urea, total bilirubin (TBil) [17], creatinine (Crea) [19],

phosphatase alkaline (ALP), aspartate aminotransferase (AST), alanine aminotransferase (ALT) [20], and some characteristics of the population (alcohol intake, smoking habits, physical activity, body mass index (BMI) [11, 23–25] have been considered.

Materials and methods

EuBIVAS population

The health status, the inclusion/exclusion criteria of the individuals enrolled in the EuBIVAS, and the used protocol have previously been described in detail [11].

To briefly summarize, EuBIVAS involved 91 presumably healthy volunteers (53 females and 38 males; age range, 21–69 years) from 6 European laboratories (Italy-Milan, Italy-Padua, Norway, Spain, the Netherlands, and Turkey). The participants completed an enrollment questionnaire to provide information about their lifestyle and presumed health status, which was further verified by a set of routine laboratory tests performed during each collection [11, 12].

Particularly, subjects were excluded from participation if any of the following criteria were met:

(1) Known diabetes and prescribed oral or insulin therapy, or fasting serum glucose >7.0 mmol/L;
(2) History of chronic liver or kidney disease;
(3) Dyslipidaemia;
(4) Family history of thalassemia syndrome and other haemoglobinopathies;
(5) Results of examinations that clearly point to a severe chronic disease (cancer, cardiovascular or neurological) or acute disease;
(6) Known carrier state for hepatitis B virus (HBV), hepatitis C virus (HCV), and human immunodeficiency virus (HIV);
(7) History of being a hospital in-patient or otherwise seriously ill
(8) Blood donation in the previous 3 months;
(9) Female subjects who were pregnant, breastfeeding, or within1 year after childbirth;
(10) Any other significant disease or disorder that, in the opinion of the investigator, could either put the subjects at risk because of participation in the study or could influence the results of the study [11].

For each of the 91 subjects, the following characteristics have been considered: age (years), sex (M/F), menopause (Yes/NO), body mass index (BMI), country of origin, smoking habits (number of cigarettes/day), physical activity (hours/week), alcohol intake (Units/day) (Table 1).

The EuBIVAS was approved by the Institutional Ethical Review Board of San Raffaele Hospital (Milan, Italy; protocol number: WG-BV project #001, 50/INT 2014) in agreement with the World Medical Association Declaration of Helsinki and by the Ethical Board/Regional Ethics Committee for each center (protocol number: WG-BV project #001, PI-1993. April 2015 for Spain; WG-BV project #001, 2014-26 for The Netherlands; WG-BV project #001, 3452/AO/15 for PD Italy; 2015-3/17 for Turkey; 2014/1988 for Norway).

Sample collection and analytical methods

Fasting serum samples were collected weekly, for 10 consecutive weeks (April–June 2015) on a set day (Tuesday to Friday), and at the same time (e.g., between 08:00 a.m. and 10:00 p.m. at each weekly visit). Sample collection was performed by the same phlebotomist at most visits, further minimizing variation. All laboratories followed the same protocol for the pre-examination phase. Serum samples were aliquoted and sent, frozen in dry ice, to the coordinating center (San Raffaele Hospital in Milan) and stored in a freezer at −80 °C until analysis (December 2017-January 2018) [11, 12].

With the exception of Alb, for the other 12 measurands here considered (Glu, Ca, TP, Na, K, Cl, urea, Crea (enzymatic method), ALP, AST, ALT, TBil), all measurements were performed with ADVIA 2400 Clinical Chemistry System (Siemens Healthineers), using Siemens reagents, calibrators, and control materials, as previously described [17, 19, 20].

Alb analyses were performed on the Roche Cobas c702 (Roche Diagnostics) using Roche reagents, and calibrated using protein-specific Roche calibrators according to the manufacturer's instructions [18]. All samples from the same participant were analyzed in duplicate within a single run. All analysis was performed at San Raffaele Hospital in Milan, Italy.

Databases

To investigate a ML approach to detect the outliers from the original EuBIVAS raw dataset, three different databases have been considered, as described below and summarized in Figure 1.

The EuBIVAS dataset included 22,280 raw data values, instead of the theoretical 23,660 (13 measurands, 10 samples in duplicate, 91 subjects): 5.8% of data were missing, since only 77 participants completed all 10 collections; while 10 participants completed nine collections, two participants completed eight collections, and further two participants completed seven collections [11, 12].

Database 1: The original database, used to publish the BV estimates EuBIVAS based, was used as the reference database or database 1 (DB1). DB1 consisted of 21, 712 results, obtained after the application of statistical tests for outlier detection (568 data values), as previously described in EuBIVAS publication [17–20].

As a first step, the missing data were imputed by means of the missing procedure (see below "*Missing data*"), to obtain the total theoretical amount of 23,660 data (91 subjects, 13 measurands, 10 collections, two replicates). As a second step, the values beyond 4 standard deviations, for each measurand distribution (at least one replicate compared to the general distribution of all subjects), were discarded. Finally, the DB1 was imputed a second time using the missing procedure.

Database 2: The original database of 22,280 EuBIVAS raw data values was used as a starting point to obtain the database 2 (DB2). As a first

Table 1: Gender, number, age, body max index (BMI), smoking habits, alcohol intake, and physical activity done by men and women <50, and women >50 years enrolled by each center.

		Men median age, years (age range)	Men median BMI, kg/m^2 (BMI range)	Physical activity			Smoking habits			Alcohol intake[a]		
				No physical activity	<7 h/ week	≥7 h/ week	0 cigarettes/ day	≥5 cigarettes/ day	>5 cigarettes/ day	0 U/ day	≤2 U/ day	>2 U/day
Italy – Milan	Men (n=9)	38 (24–59)	25.2 (20.8–30.0)	5	3	1	6	0	3	2	5	2
(19 people)	Women <50 (n=7)	34 (24–48)	22.7 (17.6–23.9)	2	2	3	6	1	0	3	4	0
	Women >50 (n=3)	58 (55–59)	22.8 (19.4–27.5)	1	1	1	3	0	0	2	1	0
Norway	Men (n=7)	37 (28–42)	24.3 (18.1–26.3)	1	2	4	6	1	0	1	3	3
(15 people)	Women <50 (n=6)	39 (29–49)	21.7 (18.7–24.4)	1	2	3	4	2	0	0	3	3
	Women >50 (n=2)	63	24.6 (23.7–25.5)	1	1	0	2	0	0	0	2	0
Spain	Men (n=7)	34 (26–54)	25.1 (19.5–32.5)	3	2	2	6	1	0	2	4	1
(16 people)	Women <50 (n=7)	26 (24–48)	21.7 (17.9–23.1)	1	5	1	5	1	1	1	6	0
	Women >50 (n=2)	60	21.3 (21.2–21.4)	0	1	1	2	0	0	0	2	0
Italy – Padua	Men (n=5)	32 (27–35)	22.5 (19.0–23.5)	1	3	1	3	2	0	3	2	0
(14 people)	Women <50 (n=8)	33 (27–49)	19.8 (18.7–23.2)	3	3	2	8	0	0	6	2	0
	Women >50 (n=1)	69	18.6	0	1	0	1	0	0	0	1	0
Turkey	Men (n=6)	27 (22–35)	27.5 (22.2–29.9)	5	1	0	4	0	2	5	1	0
(15 people)	Women <50 (n=9)	33 (21–38)	21.1 (18.3–27.3)	8	1	0	4	4	1	3	6	0
	Women >50 (n=0)	/	/	/	/	/	/	/	/	/	/	/
The Netherlands	Men (n=4)	36 (23–45)	24.0 (18.1–26.3)	1	1	2	4	0	0	1	2	1
(12 people)	Women <50 (n=6)	39 (29–49)	21.7 (20.9–24.2)	0	3	3	5	0	1	0	6	0
	Women >50 (n=2)	60 (59–60)	23.0 (20.7–25.3)	0	0	2	1	0	1	0	2	0
Total	Men (n=38)	35 (22–59)	24.4 (18.1–32.5)	16	12	10	29	4	5	14	17	7
(91 people)	Women <50 (n=43)	34 (21–49)	21.3 (17.6–27.3)	15	16	12	32	8	3	13	27	3
	Women >50 (n=10)	60 (55–69)	22.1 (18.6–27.5)	2	4	4	9	0	1	2	8	0

[a]One alcohol unit (U) corresponds to 10 mL, equivalent to 8 g, of pure alcohol (https://www.drinkaware.co.uk/alcohol-facts/alcoholic-drinks-units/what-is-an-alcohol-unit/).

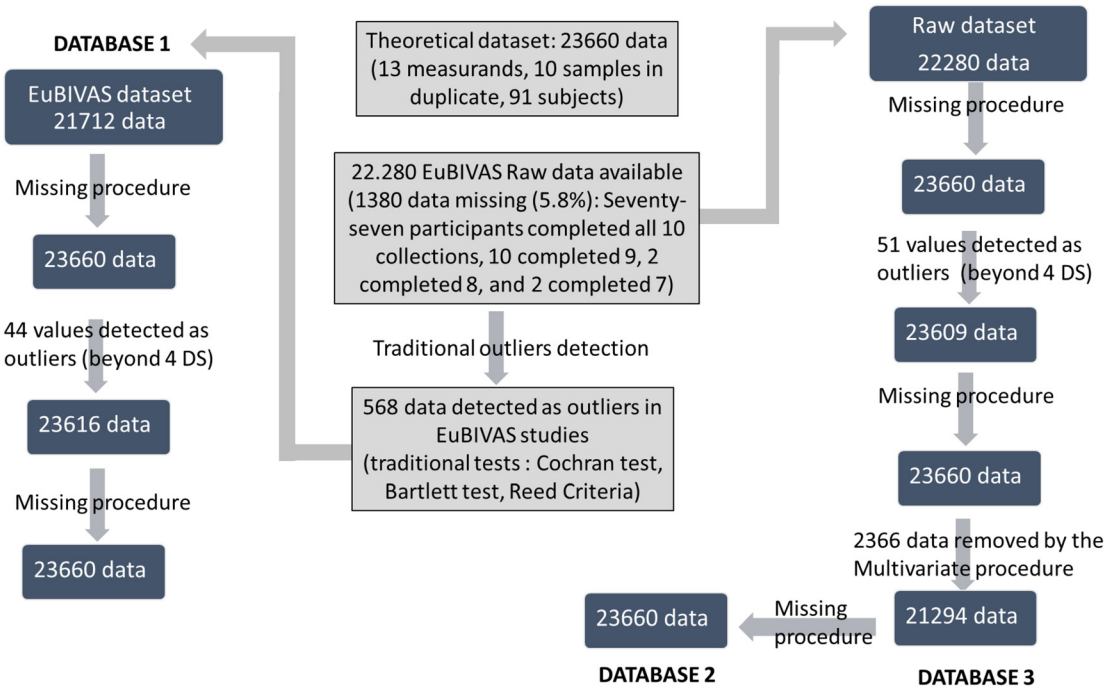

Figure 1: a flow chart describes the three databases (DB1, DB2 and DB3).
The flow chart describes the process to obtain three databases: the initial number of results, the number of outliers detected, the number of missing values replaced and the final number of data considered for database 1 (DB1), database 2 (DB2) and database 3 (DB3) respectively.

step, the missing data were imputed, to obtain the total theoretical amount of 23,660 data. As a second step, the values beyond 4SD, for each measurand distribution, were discarded. As a third step, the missing data was imputed a second time using the missing procedure (Figure 1).

Subsequently, a multivariate outlier detection technique, called Elliptic Envelope [26], was applied. Elliptic Envelope is based on the robust covariance estimation of the data. This method assumes that data is sampled from a multivariate Gaussian distribution (which can be visualized in the multi-dimensional space as an ellipse), so that outlier observations are defined as those observations lying "too far" outside of the ellipse fitted around the data. The outlier-ness of the observations is determined through the *Mahalanobis distance*. Briefly, the Mahalanobis distance is the distance between a multidimensional point and a distribution, computed as $d(x) = [(x-\mu)^T \cdot C^{-1} \cdot (x-\mu)]^{0.5}$, where x is the vector of the observations, μ the vector of mean values, T denotes matrix transposition, and C^{-1} the inverse covariance matrix [22, 26]. Compared with traditional univariate outlier detection approaches, this method, being based on a multivariate criterion, takes into account, at the same time, all results obtained from the same sample (two replicates for 13 measurands). Therefore, for each outlier detected 26 results were discarded (one sample). As a final step, the DB2 was imputed a second time, using the missing procedure, to obtain the total theoretical amount of 23,660 data (Figure 1).

Database 3: The last database, database 3 (DB3), is the most conservative database. It differs from the DB2 in that the last missing procedure is not performed. By contrast, missing results are discarded from further analysis. Thus, the final amount of data included in DB3 is lower than that included in DB1 and DB2, and it corresponds to the data obtained after having applied the Elliptic Envelope ML technique (Figure 1).

Missing data

The missing value imputation has been applied as follows:

(1) If a subject, for a given collection of a particular measurand, has only one of the two replicates, then the missing replicate is imputed with the mean value of the replicate ± SD of the same subject (+SD if the value of the present replicate is lower than the mean, and −SD if higher);
(2) If both replicates for a given collection of an analyte were missing, then the missing values are imputed with two values having the same mean value and SD obtained from the other samples of the same subject;
(3) If a subject does not have any result for a given analyte, then its 20 results are imputed with 20 values sampled at random from a distribution with the same mean values and SD than the other subjects from the same population subgroup (males, fertile females or menopausal females).

Data matrices

From the resulting databases (DB1, DB2 and DB3), for each subject for each analyte, the mean values were calculated, aggregating the ten (or all the available ones, in the case of the DB3) remaining couples of

replicates. Therefore, only one value for each analyte is assigned to each subject.

For each database, the first data matrix (DM-mean) considered encompassed 91 rows (subjects) and 15 columns (features) reporting sex, age and quantitative variables: 13 mean values, one for each measurand.

A second data matrix (DM-CV_P) has been considered, using, for each single subject, the "within-person BV" (CV_P), for each measurand, instead of the mean values. The DM-CV_P is therefore composed by 91 rows (subjects) and 13 CV_P values, one for each measurand.

To note that the term CV_P is used instead of the "within-subject BV (CV_I)" to distinguish between the coefficients of variation for a single individual (CV_P) and for the pooled value from studies on a number of individuals (CV_I), respectively, as recently by Simundic et al. [27].

CV_P, for each subject and for each measurand, has been obtained from the 20 results (2 replicates for 10 collections) by computing the CV_A within person (CV_{AP}) between replicates using the following formula:

$$DS_P=(\sigma_P^2-\sigma_A^2)^{\wedge 1/2}$$
$$CV_P=(DS_P)/X_P{}^*100$$

Where:
- σ^2_P is the total biological variance for a single person
- σ^2_A is the analytical variance for a single person
- DS_P is intra person biological standard deviation
- X_P is the mean value for a single person

The 2 principal component analysis (2PCA)

PCA is a machine learning technique that has the goal of performing dimensionality reduction, while preserving most of the information about the variance within data. As a consequence, this method is particularly useful to visualize multidimensional data.

Briefly, PCA transform the data from a 13-dimensional space, mapping them into a k-dimensional space (kPCA). In this work, we selected a value of k = 2, ass this technique was used to visualize whether there may be strongly recognizable clusterings.

The 2PCA representations of DM-mean and DM-CV_P were plotted, so as to visually inspect any possible clustering among subjects. Each point in the picture represents a single subject in a two-dimensional space, obtained from the original 13 features through the application of 2PCA.

The demographic characteristics reported in Table 1 have been evaluated, as possible clusters, for each database, in both DM-mean and DM-CV_P, as follows:
- Three gender/age related subgroups: males, fertile females and menopausal females;
- Five subgroups related to the country of origin: Italy, Norway, Spain, The Netherland and Turkey;
- Three physical activity subgroups: none, light(<7 h/week), and strong physical activity (>7 h/week);
- Two BMI subgroups: <21 and ≥21;
- Three smoking habits subgroups: none; ≤5, >5 cigarettes/day;
- Three alcohol intake subgroups: none; moderate (≤2U/day), high (>2Unit day);

The data elaboration has been performed using python 3.7, with the following libraries: pandas version 1.2.0; matplotlib version 3.3.2; numpy version 1.19.2; seaborn version 0.11.1; sklearn version 0.23.2; scipy version 1.5.2. Minibatch K-means and Spectral Clustering algorithms have been used to estimate the accuracy of the clusterization procedure. In particular, accuracy was estimated, for each of the above mentioned possible demographic characteristics, by matching each cluster with the population group for which the degree of overlap was maximal.

Results

DB1. The number of outliers detected and removed by the traditional statistical analysis for the DB1 was previously published [17–20], while the number of the imputed missing values, together with the number of the second outlier detection procedure, are reported in Figure 1.

DB2 and DB3. The flowchart (Figure 1) reports the number of the imputed missing values, the number of observations detected as outliers beyond the 4SD, and the number of observations (samples) detected as outliers by the Elliptic Envelope technique (multivariate procedure), for both DB1 and DB2. Fifty-one observations were detected as outliers as beyond the 4SD from the general distribution, while the assessment of the Elliptic Envelope ML technique led to the exclusion of further 91 samples (10% of the total set of data). These 91 samples have been removed from 49 subjects, resulting in a mean of 1.9 observations/subject.

A subject without any data for a given measurands never occurred, so that the imputation of 20 values randomly generated with the same mean values and DS of the other subjects was not necessary.

For each database, the total number of imputed missing values, for each measurand, is reported in Table 2.

The PCA visualizations obtained for the three databases (panel a, b and c), for the mean and CV_P values, are reported in Figures 2 and 3 respectively. The subjects are represented through different colors, according to the gender/age groups. A clear clustering has been identified between males and females mean values (Figure 2) with an accuracy of 86, 86 and 85% for DB1, DB2, and DB3 respectively, while for CV_P values no clustering has been found (accuracy around 60% in all databases) (Figure 3). Interestingly, the points that identify the mean values of the females in menopausal age, are plotted closer to the male subgroup. Data interpretations derived from the three different databases are similar. The feature importance scores, computed by means of the PCA analysis, for the three different databases for the mean values clustering are shown in supplemental figure 1. ALP, followed

Table 2: Number of data included for each database for the measurands considered.

Measurand	Database 1		Database 2		Database 3	
	Initial number of data	Final number, % of missing replaced	Initial number of data	Final number, % of missing replaced	Initial number of data	Final number, % of missing replaced
GLU	1,658	1,820, 9.0%	1,713	1,820, 16.2%	1,713	1,638, 6.9%
Ca	1,700	1,820, 6.6%	1,712	1,820, 16.0%	1,712	1,638, 6.7%
Alb	1,675	1,820, 8.0%	1,745	1,820, 14.2%	1,745	1,638, 4.6%
TP	1,698	1,820, 6.7%	1,713	1,820, 15.9%	1,713	1,638, 6.5%
Na	1,664	1,820, 8.6%	1,718	1,820, 15.7%	1,718	1,638, 6.3%
K	1,706	1,820, 6.3%	1,716	1,820, 15.7%	1,716	1,638, 6.3%
Cl	1,713	1,820, 5.9%	1,718	1,820, 15.9%	1,718	1,638, 6.5%
Urea	1,660	1,820, 8.8%	1,712	1,820, 16.0%	1,712	1,638, 6.7%
Crea	1,716	1,820, 5.7%	1,716	1,820, 15.7%	1,716	1,638, 6.3%
ALP	1,618	1,820, 12.1%	1,714	1,820, 16.5%	1,714	1,638, 7.3%
AST	1,655	1,820, 9.1%	1,699	1,820, 17.1%	1,699	1,638, 7.9%
ALT	1,566	1,820, 14.7%	1,702	1,820, 16.6%	1,702	1,638, 7.3%
TBil	1,683	1,820, 8.1%	1,702	1,820, 17.1%	1,702	1,638, 7.9%
Total	21,712	23,660, 8.4%	22,280	23,660, 16.0%	22,280	21,294, 6.7%

GLU, glucose; Alb, albumin; TP, total proteins; Na, sodium; K, potassium; Cl, chloride; Crea, creatinine; ALP, phosphatase alkaline; AST, aspartate aminotransferase; ALT, alanine aminotransferase; TBil, total bilirubin.

by Crea, ALT and AST, resulted as the most significant features for all databases.

As for the gender/age subgroups, no significant differences have been found, in the three databases, also for country of origin, smoking habits, alcohol intake, BMI and physical activity (data not shown). For these latter, PCA visualization are shown only for the DB3.

No clustering related to the country of origin was detectable (Figure 4), for both mean values (Figure 4A), and CV_PS values (Figure 4B). Figure 5 reports the PCAs for the physical activity (panels a, b) and for the BMI (panels c, d). A negligible clustering (accuracy 63%) appears for the mean values according to the BMI (panel c). Similarly, irrelevant clusterings for smoking habits, and alcohol intake are shown in Figure 6, for both mean and CV_P values.

Discussion

Although ML is routinely used in biochemical development and for evaluating and interpreting data in several medical specialties, the application of ML in clinical laboratory medicine still appears to be an unexpectedly slow process [28]. However, "A short guide for medical professionals in the era of artificial intelligence" has recently been

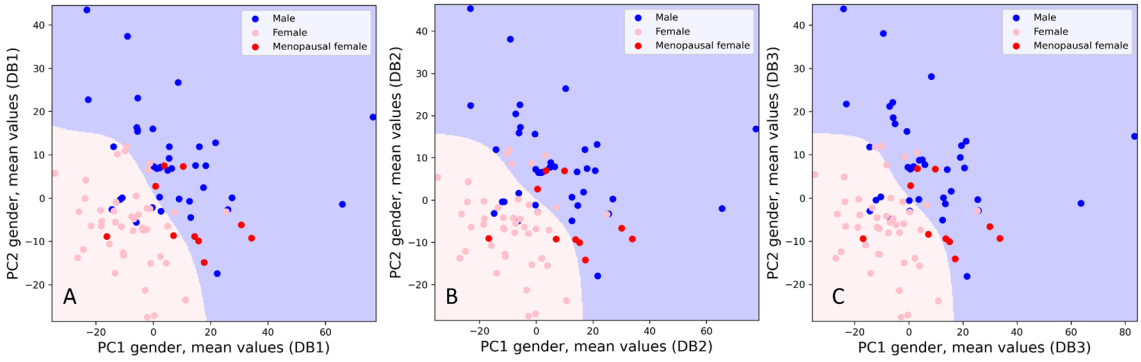

Figure 2: Mean values gender related 2 principal component analysis (2PCA).
Each point in the picture represents a single subject position in a space in a two dimensional representation (2PCA) obtained using 13 mean values as features. The blue, pink and red circles indicate the men, females in fertile and in menopausal age respectively. The data from database 1 (DB1), database 2 (DB2), and database 3 (DB3), are represented in the three different panels (A, B and C, respectively).

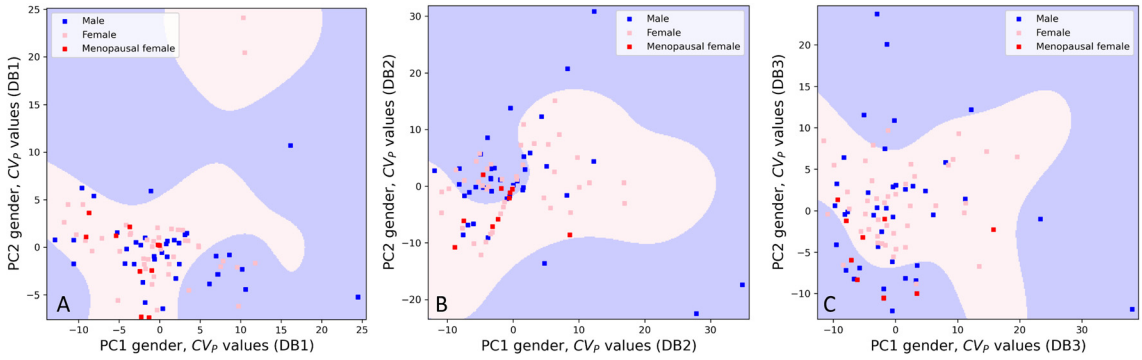

Figure 3: CV_P values gender related 2 principal component analysis (2PCA).
Each point in the picture represents a single subject position in a space in a two dimensional representation (2PCA) obtained using 13 CV_P values as features. The blue, pink and red circles indicate the men, females in fertile and in menopausal age respectively. The data from database 1 (DB1), database 2 (DB2), and database 3 (DB3), are represented in the three different panels (A, B and C, respectively).

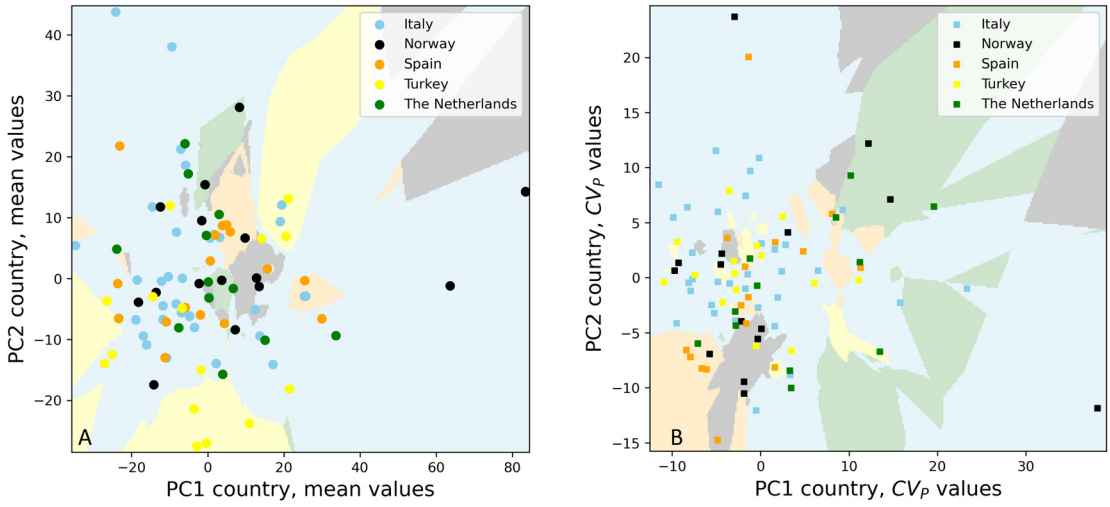

Figure 4: Mean and CV_P values countries related 2 principal component analysis (2PCA).
Each point in the picture represents a single subject position in a space in a two dimensional representation (2PCA). The different colors represent the country of origin: light blue for Italy, black for Norway, orange for Spain, yellow for Turkey, and green for The Netherland. The mean values and CV_P data are represented in panels (A) and (B), respectively.

published, with the aim to serve as a short repository of information and details every physician might need to know in the age of AI [29]. Laboratory data are influenced by several factors including the pre-analytical (collection, time, stability, storage temperature, etc.) [30], and the analytical phase (standardization, harmonization etc.) [31, 32], therefore they should be used with caution and competence, in collaboration with clinicians [21].

Hereby, we applied ML techniques on a metabolic panel of measurands from the EUBIVAS study, with a dual purpose: first, to verify the homogeneity of the involved population, coming from different European countries; second, to evaluate the existence of possible clusterings, due to different lifestyles, and to validate a new multivariate approach to detect outliers using ML techniques. For this purpose, the EuBIVAS database, used in previously publication [17–20], cleaned of the outliers detected by the traditional univariate statistical tests, was used, and compared to other two databases obtained as described in the previous sections. Furthermore, since most ML algorithms cannot be applied on datasets with missing data, a missing data imputation procedure was applied. In this respect, it is worth noting that the missing procedure here conceived, has been established to do not alter the mean and variance values either in the single subject, or in the subgroup, or in the whole population.

The number of data removed by the multivariate procedure to obtain the databases 2 and 3 (Figure 1) is of

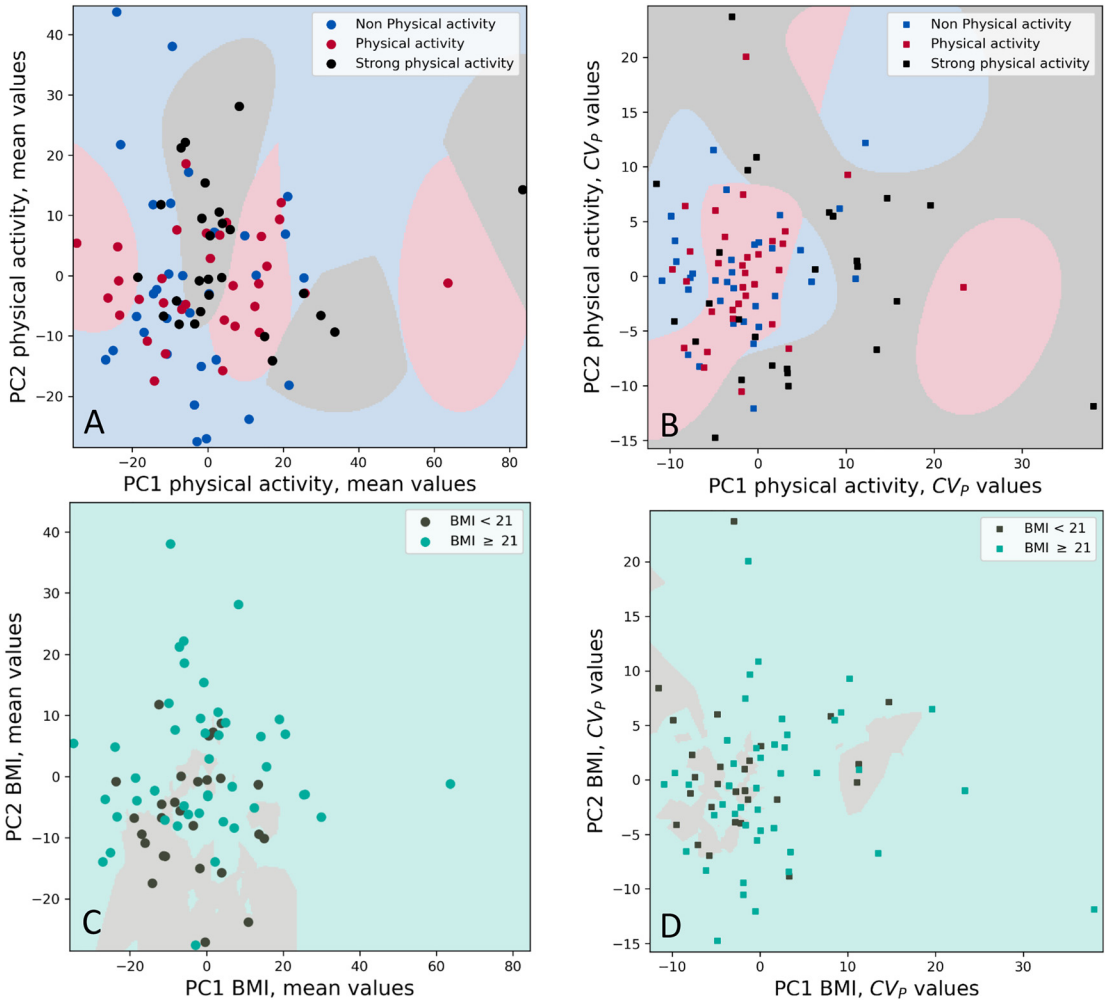

Figure 5: Mean and CV_P values according to physical activity and BMI 2 principal component analysis (2PCA).
Each point in the picture represents a single subject position in a space in a two dimensional representation (2PCA). Panels (A) and (B): different colors represent different amounts of physical activity: none in blue, medium in red (<7 h/week), and strenuous in black (>7 h/week) for mean values and CV_P data (panel a and b respectively). Panels (C) and (D): body mass index (BMI) < 21 and BMI > 21 for mean and CVP values are represented in black and in light blue in (C) and (D), respectively.

2,366 results (10% of the total number of data), much higher than the number of data (568 data, 2.4%) removed by the traditional statistical tests used individually for each measurand. This difference in the number of outliers identified and removed should not surprise, if we considered that multivariate methods can identify observations which deviate from the global behavior of the data (that is, observations that have relationships between variables which are different from most of the other observations), even in the absence of extreme values in each single variable (univariate approach).

Despite the different methods applied to detect the outliers, and the differences among the number of data removed, the visualizations obtained applying PCA analysis to the three databases were similar: this confirms the face validity of the ML multivariate method used to remove the outliers, and of the missing replacement procedure.

Figure 2 shows an evident clustering in all databases, based on the mean values for males and females subgroups. This clustering, here identified, means that with an accuracy of 85% it is possible to predict, from laboratory data, whether the subject is male or female. Such a relevant clusterization, using mean values of the measurands here considered, should not be considered surprising as it is obviously due to the different values concentration between males and females for some measurands, which is also reflected in different reference intervals (ALP, creatinine, ALT and AST). Interestingly, these measurands were identified as the most important features for the PCA

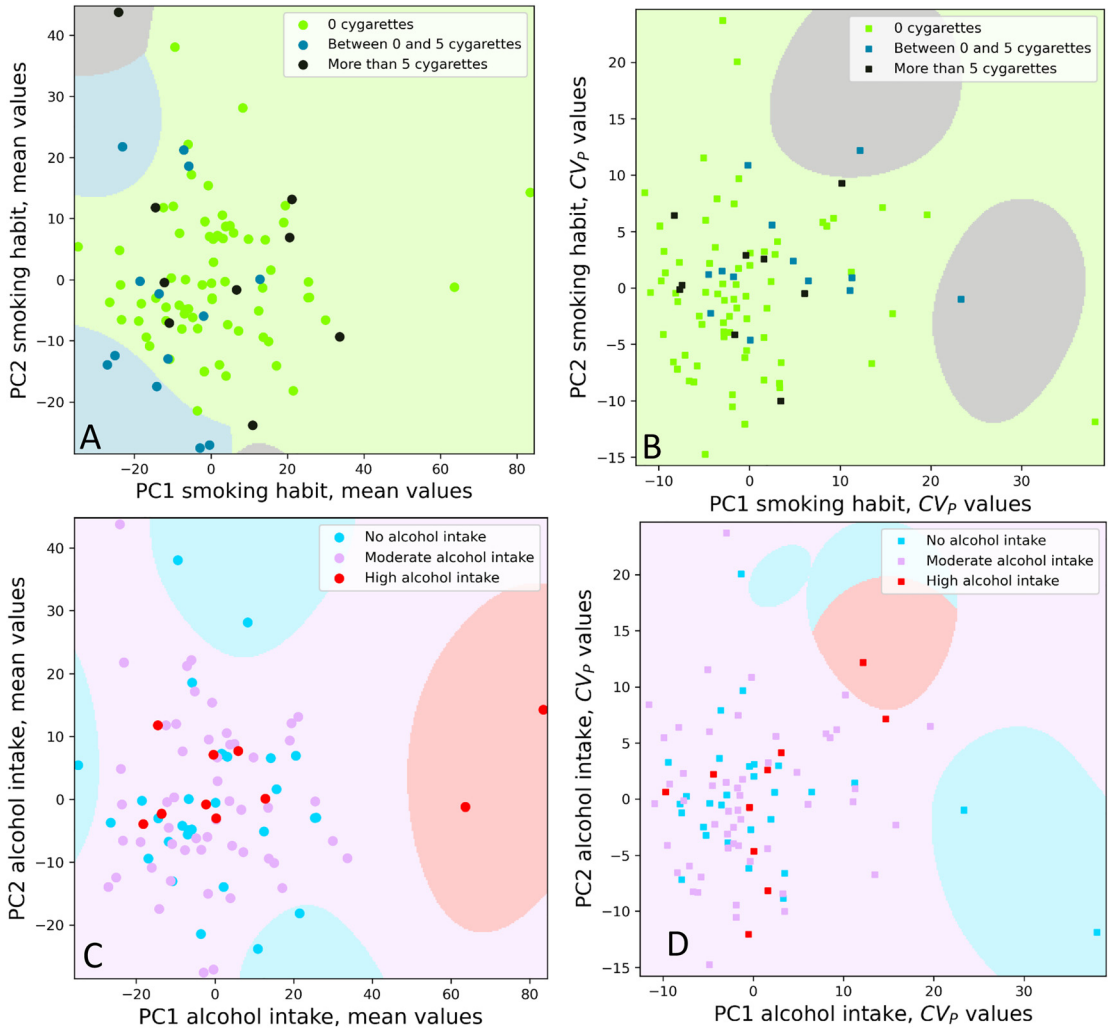

Figure 6: Mean and CV_P values according to alcohol consumption and smoking habits 2 principal component analysis (2PCA). Each point in the picture represents a single subject position in a space in a two dimensional representation (2PCA). Panels (A) and (B): different colors represent different smoking habits: none in green, low in blue (<5 cigarettes/day), and high in black (>5 cigarettes/day) for mean values and CV_P data (panel a and b respectively). Panels (C) and (D): different colors represent different alcohol intake: none in light blue, moderate in pink (<2 Units/day), and high in red (>2 Units/day) for mean values and CV_P data (panels (A) and (B), respectively).

clustering (supplemental figure 1). However, as shown in Figure 2, it is interesting to note that the same clustering has been found in all databases with almost the same accuracy (between 85 and 86%), confirming the fact that the databases are able to provide the same information, despite different approaches in missing values identification and replacement. This further suggests that observations identified as outliers by the multivariate method were also labelled as outliers by the univariate method. Moreover, this expected clustering with gender indirectly confirms that the PCA approach works. Another noteworthy finding, regards the subgroup of menopausal females that are overall displaced from the female area into male area, indicating the need to set appropriate reference intervals for females in menopausal age, for which more studies are desirable. The six males that are moved in females area (mean age 34.8 ± 3.3 years), three from Italy and three from Turkey, are characterized to be sedentary (5 out 6 declared to do not do any physical activity, and just one only light activity), to be drinkers (6 out of 5 more than 2U/day of alcohol consumption) and no smokers (just one out of 6 declared to smoke), and to have a high BMI (mean 26.3 ± 2.6).

Contrary to that observed for the mean values, the CV_P values obtained in males' and females' subgroups,

do not show any visible clustering in any database (Figure 3). All this seems to suggest that, despite the differences in mean values, the component of BV is not overall affected by the concentration gender-related, confirming the validity in using a unique RCV and APS values based on BV estimates.

Considering that the three databases have always given overlapping results (data not shown), for convenience, PCAs related to the data of the only DB3 are shown in the next figures.

The countries of origin of the subjects enrolled in EuBIVAS are highlighted by different colors in Figure 4 for mean (a) and CV_P values (b) respectively. In Figure 4, no clustering for the country is recognizable, which implies that it is not possible to distinguish any sub-population from their origin and, as a consequence, the whole EUBIVAS population can be considerable homogenous. Actually, EuBIVAS studied previously published about the measurands here included, found a significant difference in population only in the case of creatinine mean values for the Turkish people [19], while for other measurands no differences among countries were found.

The observation that the mean creatinine concentration was lower in Turkish participants is not surprising. This has been reported previously, being explained as resulting from differences in diet (low meat consumption) and different physical activity [19]. This was also confirmed by the lifestyle information provided by the enrolled subjects in EuBIVAS (Table 1). In this study, the small difference in creatinine concentration, is "hidden" in a multi-measurand panel, so that the EuBIVAS population all together can be considerate suitable, being homogenous, for having estimated BV generizable data. This last consideration further stresses the importance that both the multivariate as well as the univariate approach may give valuable information.

From the PCAs visualization (Figure 4), there is no data suggesting any differences in pre-analytical variables or samples treatment among the different laboratories. This finding is particularly relevant, considering the criticality in the pre-analytical phase of some measurands included in the metabolic panel. For example, several pre-analytical quality indicators in laboratory medicine are strongly dependent on the time the sample is collected, and careful attention to the pre-analytical phase is essential to ensure accurate glucose and electrolytes measurements for the delay in specimen processing [33–35]. Moreover, many factors can affect analyte stability for most laboratory tests. Just think about the enzyme activities measurements, where sampling time and storage information are crucial for qualifying a sample as suitable for being tested [33, 36].

Conclusions

The EuBIVAS protocol followed a strict experimental design [11, 12], powered to deliver estimates of BV with a high degree of reliability for a well-characterized multinational cohort of subjects.

The absence of meaningful differences between groups from Turkey, Norway, The Netherlands, Spain, and Italy confirms that the obtained data are widely applicable across healthcare systems and that they can be used to deliver APS for systems to be used internationally.

Similarly to what has been observed for the country variable, the absence of clusterings for alcohol, smoking habits, BMI and physical activity further confirms that the EuBIVAS sample is homogenous.

In conclusion, our data support the use of ML PCA technique and of the multivariate approach to detect outliers, in alternative to univariate methods, to gain new insight of the data. Moreover, this novel approach confirms the homogeneity of the EuBIVAS dataset.

Acknowledgments: The authors would like to thank the EFLM Working Group on BV for the use of data from the EuBIVAS.
Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Informed consent was obtained from all individuals included in this study.
Ethical approval: The EuBIVAS protocol was approved by the Institutional Ethical Review board of San Raffaele Hospital in agreement with the World Medical Association Declaration of Helsinki and by the Ethical board/regional Ethics Committee for each involved center. Informed consent was signed by all enrolled subjects.

References

1. Fraser CG, Kallner A, Kenny D, Petersen PH. Introduction: strategies to set global quality specifications in laboratory medicine. Scand J Clin Lab Invest 1999;59:477–8.

2. Haeckel R, Wosniok W, Kratochvila J, Carobene A. A pragmatic proposal for permissible limits in external quality assessment schemes with a compromise between biological variation and the state of the art. Clin Chem Lab Med 2012;50:833–9.
3. Carobene A, Franzini C, Ceriotti F. Comparison of the results from two different External Quality Assessment Schemes supports the utility of robust quality specifications. Clin Chem Lab Med 2011; 49:1143–9.
4. Fraser CG. Reference change values: the way forward in monitoring. Ann Clin Biochem 2009;46:264–5.
5. Fraser CG. The nature of biological variation. In: biological variation: from principles to practice. Washington, DC: AACC Press; 2001. pp. 1–27.
6. Coskun A, Sandberg S, Unsal I, Cavusoglu C, Serteser M, Kilercik M, et al. Personalized reference intervals in laboratory medicine: a new model based on within-subjects biological variation. Clin Chem 2021;67:374–84.
7. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. Clin Chem Lab Med 2015;53:829–32.
8. Sandberg S, Fraser GC, 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.
9. Carobene A. Reliability of biological variation data available in an online database: need for improvement. Clin Chem Lab Med 2015; 53:871–7.
10. Aarsand AK, Røraas T, Bartlett WA, Coşkun A, Carobene A, Fernandez-Calle P, et al. Harmonization initiatives in the generation, reporting and application of biological variation data. Clin Chem Lab Med 2018;56:1629–36.
11. Carobene A, Strollo M, Jonker N, Barla G, Bartlett WA, Sandberg S, et al. Sample collections from healthy volunteers for biological variation estimates' update: a new project undertaken by the Working Group on Biological Variation established by the European Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem Lab Med 2016;54:1599–608.
12. Carobene A, Aarsand AK, Bartlett WA, Coskun A, Diaz-Garzon J, Fernandez-Calle P, et al. The European biological variation study (EuBIVAS): a summary report. Clin Chem Lab Med 2021. https://doi.org/10.1515/cclm-2021-0370 [Epub ahead of print].
13. Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. Clin Chem 2012;58:1306–13.
14. Carobene A. The European biological variation study (EuBIVAS): delivery of updated biological variation estimates, a project by the working group on biological variation in the European federation of clinical Chemistry and laboratory medicine. J Lab Precis Med 2017;2:70.
15. Aarsand A, Roraas T, Fernandez-Calle P, Ricós C, Diaz-Garzon J, Jonker N, et al. On behalf of the EFLM Working Group on Biological Variation and Task and Finish Group for the Biological Variation Database. The biological variation data critical appraisal checklist (BIVAC): a new standard for evaluating studies on biological variation. Clin Chem 2018;64:501–14.
16. Bartlett WA, Braga F, Carobene A, Coşkun A, Prusa R, Fernandez-Calle P, et al. Biological variation working group, European federation of clinical Chemistry and laboratory medicine (EFLM). A checklist for critical appraisal of studies of biological variation. Clin Chem Lab Med 2015;53:879–85.
17. Aarsand AK, Díaz-Garzón J, Fernandez-Calle P, Guerra E, Locatelli M, Bartlett WA, et al. The EuBIVAS: within- and between-subject biological variation data for electrolytes, lipids, urea, uric acid, total protein, total bilirubin, direct bilirubin, and glucose. Clin Chem 2018;64:1380–93.
18. Carobene A, Aarsand AK, Guerra E, Bartlett WA, Coskun A, Díaz-Garzón Marco J, et al. European biological variation study (EuBIVAS): within- and between-subject biological variation data for 15 frequently measured proteins. Clin Chem 2019;65:1031–41.
19. Carobene A, Marino I, Coŝkun A, Serteser M, Unsal I, Guerra E, et al. The EuBIVAS project: within and between-subject biological variation data for serum creatinine using enzymatic and alkaline picrate methods and implications for monitoring. Clin Chem 2017;63:1527–36.
20. Carobene A, Røraas T, Sølvik UØ, Sylte MS, Sandberg S, Guerra E, et al. Biological variation estimates obtained from 91 healthy study participants for 9 enzymes in serum. Clin Chem 2017;63: 1141–50.
21. Badrick T, Banfi G, Bietenbeck A, Cervinski MA, Loh TP, Sikaris K. Machine learning for clinical chemists. Clin Chem 2019;65:1350–6.
22. Ghorbani H. Mahalanobis distance and its application for detecting multivariate outliers. Facta Univ – Ser Math Inf 2019;34: 583–95.
23. Bottani A, Banfi G, Locatelli M, Aarsand AK, Coşkun A, Díaz-GarzónJ, et al. European biological variation study (EuBIVAS): within- and between-subject biological variation estimates for serum thyroid biomarkers based on weekly samplings from 91 healthy participants. Clin Chem Lab Med 2021. https://doi.org/10.1515/cclm-2020-1885 [Epub ahead-of-print].
24. Bottani M, Banfi G, Guerra E, Locatelli M, Aarsand AK, Coşkun A, et al. European Biological Variation Study (EuBIVAS): within- and between-subject biological variation estimates for serum biointact parathyroid hormone based on weekly samplings from 91 healthy participants. Ann Transl Med 2020;8:855.
25. Cavalier E, Fraser CG, Bhattoa HP, Heijboer AC, Makris K, Ulmer CZ, et al. Analytical performance specifications for 25-hydroxyvitamin D examinations. Nutrients 2021, 13, 431. doi.org/https://doi.org/10.3390/nu13020431.
26. McKinnon C, Carroll J, McDonald A, Koukoura S, Infield D, Soraghan C. Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data. Energies 2020;13:5152.
27. Simundic AM, Kackov S, Miler M, Fraser CG, Petersen PH. Terms and symbols used in studies on biological variation: the need for harmonization. Clin Chem 2015;61:438–9.
28. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med 2018;56:516–24.
29. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126.
30. Vermeersch P, Frans G, von Meyer A, Costelloe S, Lippi G, Simundic AM. How to meet ISO15189:2012 pre-analytical

requirements in clinical laboratories? A consensus document by the EFLM WG-PRE. Clin Chem Lab Med 2021;59:1047–61.
31. Miller WG, Greenberg N. Harmonization and standardization: where are we now? J Appl Lab Med 2021;6:510–21.
32. Carobene A, Ceriotti F, Infusino I, Frusciante E, Panteghini M. Evaluation of the impact of standardization process on the quality of serum creatinine determination in Italian laboratories. Clin Chim Acta 2014;427:100–6.
33. Lippi G, Betsou F, Cadamuro J, Cornes M, Fleischhacker M, Fruekilde P, et al. Simundic AM; working group for preanalytical phase (WG-PRE), European federation of clinical Chemistry and laboratory medicine (EFLM). Preanalytical challenges - time for solutions. Clin Chem Lab Med 2019;57:974–81.
34. Janssen K, Delanghe J. Importance of the pre-analytical phase in blood glucose analysis. Acta Clin Belg 2010;65:311–8.
35. Baruah A, Goyal P, Sinha S, Ramesh KL, Datta R. Delay in specimen processing-major source of preanalytical variation in serum electrolytes. J Clin Diagn Res 2014;8:CC01–3.
36. Cuccherini B, Nussbaum SJ, Seeff LB, Lukacs L, Zimmerman HJ. Stability of aspartate aminotransferase and alanine aminotransferase activities. J Lab Clin Med 1983; 102:370–6.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2021-0599).