The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine

eJIFCC

Basics of laboratory statistics

Vivek Pant, Santosh Pradhan, Keyoor Gautam

Samyak Diagnostic Pvt Ltd, Bench to Clinic Research Center, Kathmandu, Nepal

ARTICLE INFO

Corresponding author:
Dr. Vivek Pant
Consultant Biochemist
Samyak Diagnostic Pvt Ltd
Bench to Clinic Research Center
Kathmandu
Nepal
Phone: +977 9841486789
E-mail: drvpant@gmail.com

Key words:
standard deviation, coefficient of variation, quality control, verification, reference interval

ABSTRACT

The strict monitoring of examinations and evaluation of newer methods or instruments is a daily routine in clinical laboratory. The automated analyzers accumulate an enormous amount of data from patients' examinations and quality control procedures. This laboratory data is meaningless if it does not generate the information that we can extend to the population of our interest. In an analytical work, the most important operation is the comparison of data, to quantify accuracy and precision and to generate meaningful explanation for clinician and patients queries. Most of the information needed in the regular laboratory work can be obtained with the use of simple convenient statistical tools. This article describes the basics of laboratory statistics, the knowledge of which answers about the application of quality control in laboratory, accuracy and diagnostic power of our examinations, variability in reports, comparison of different methods and derivation of a biological reference interval for an analyte.

INTRODUCTION

In the clinical laboratory, statistics are used to verify and monitor the performance of analytical methods and to guide the clinical interpretation of laboratory data. Laboratory statistics can be broadly described under following headings

1. Quality control and statistics
2. Diagnostic power of a laboratory test
3. Variability in Reports
4. Method Comparison
5. Reference Interval

QUALITY CONTROL

Quality control is the analysis of control materials, comparing the results with a predefined acceptable limit and plotting a result in a chart. Internal quality control data is best visualized using a Levey-Jennings control chart where the dates of analyses are plotted along X-axis and control values are plotted on Y-axis. The mean and one, two and three standard deviation (SD) limits are also marked on the Y-axis. Inspecting the pattern of plotted points provides a simple way to detect random error and shifts or trends in the calibration. Daily repeating the same control sample should produce a normally distributed set of data. This means, approximately 66% of values should fall between +/-1SD ranges and be evenly distributed on either side of mean. Similarly, 95% and 99 % of values should fall within +/-2SD and +/-3SD limits respectively. (Figure 1) A calculation of mean, standard deviation and coefficient of variation (CV) of this

Figure 1 Normal distribution curve

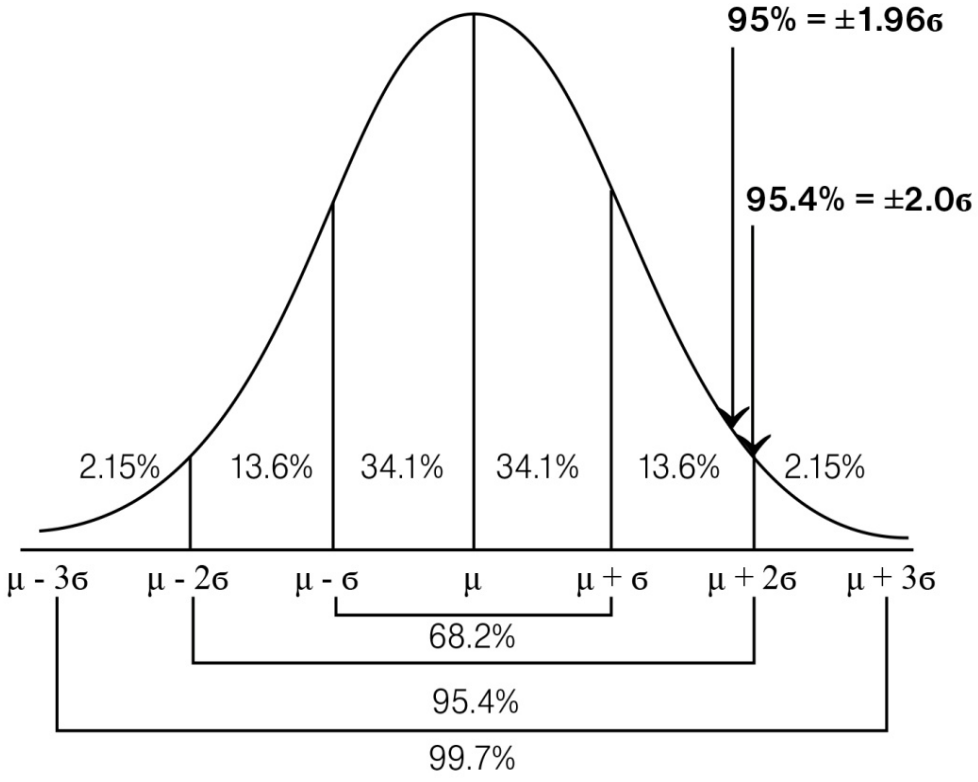
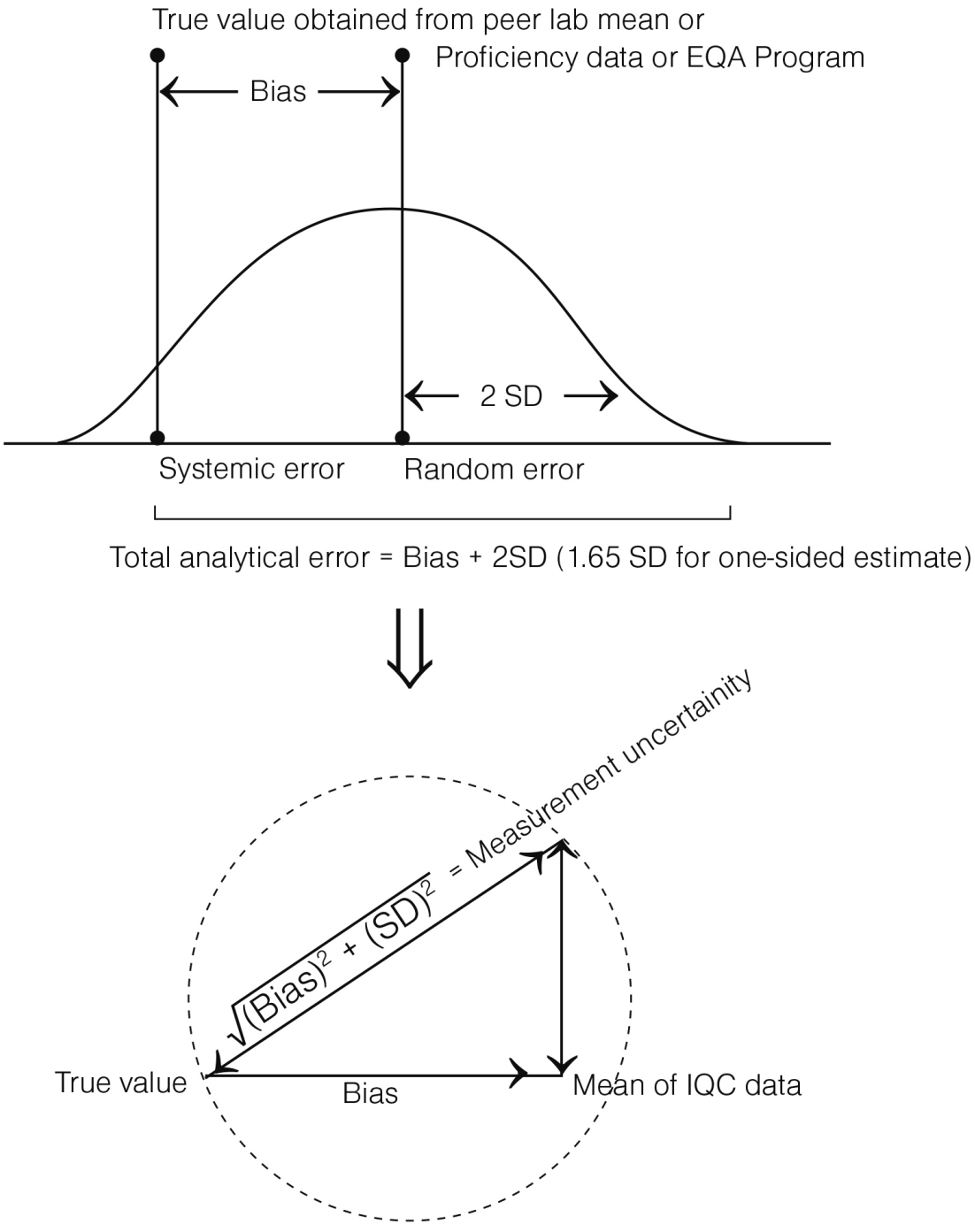


Figure 2 **Illustration of systemic and random error and concept of measurement of uncertainty**



dataset is useful for further calculation and derivation of other laboratory entities (Figure 2). Mean is the average value of measurements and SD is the primary measure of dispersion or variation of the individual result from the mean value. To derive SD, we calculate the deviation from the mean for each observation; square those results, sum them, divide by the number of observations minus one, and finally take a square root. CV is the SD expressed as a percent of the mean. Acceptable CV needs to be defined for each analyte based on medical significance.

Quality control rules are designed to detect two types of error, systemic error or bias and random error or imprecision. Precision is the agreement with replicate measurements and therefore the imprecision is caused by increased random error. Accuracy is the agreement between best estimate of the mean of results and its true value, therefore inaccuracy is caused by increased systemic error. These two errors when combined give a total analytical error (Figure 3). [1] In practice, replicate measurements can reduce, but not completely eliminate systematic and random errors, and therefore total error cannot be exactly known. [2] It follows that the true value of a measured quantity cannot be exactly known either. This assumption is fundamental to the measurement of uncertainty (MU) approach.

MU approach focuses on identifying the dispersion of results that might have been obtained for an analyte if a sample had been measured repeatedly. To do this, the MU approach uses available data about repeated measurements from a given measuring system to define an interval of values within which the true value of the measured analyte is believed to lie, with a stated level of confidence. This can be simply estimated from the CV calculated from repeated measurements of internal quality control sample. (Figure 2 and 3) In the MU concept, a measurement result can comprise two uncertainties (i) that associated with a bias correction, and (ii) the uncertainty due to random effects. [3] Both these uncertainties are expressed as SDs which, when combined together, provide the combined standard uncertainty for the procedure. (Figure 2 and 3)

External quality control (EQC) refers to the process of controlling the accuracy of an analytical method by interlaboratory comparisons. Two of the most important comparison statistics of an Interlaboratory program are the coefficient of variation ratio (CVR) and standard deviation index (SDI), which are consensus-based metrics of imprecision and bias, respectively.

The CVR allows evaluating imprecision relative to the consensus group and is expressed mathematically by the formula: Lab CV/ Consensus group CV

If the labs imprecision is equal to the imprecision of consensus group, then CVR will be 1.0.

The SDI or Z-score is a useful parameter for evaluating bias relative to the consensus group and is expressed mathematically by the formula:

(Lab mean-Consensus group mean)/ Consensus group SD

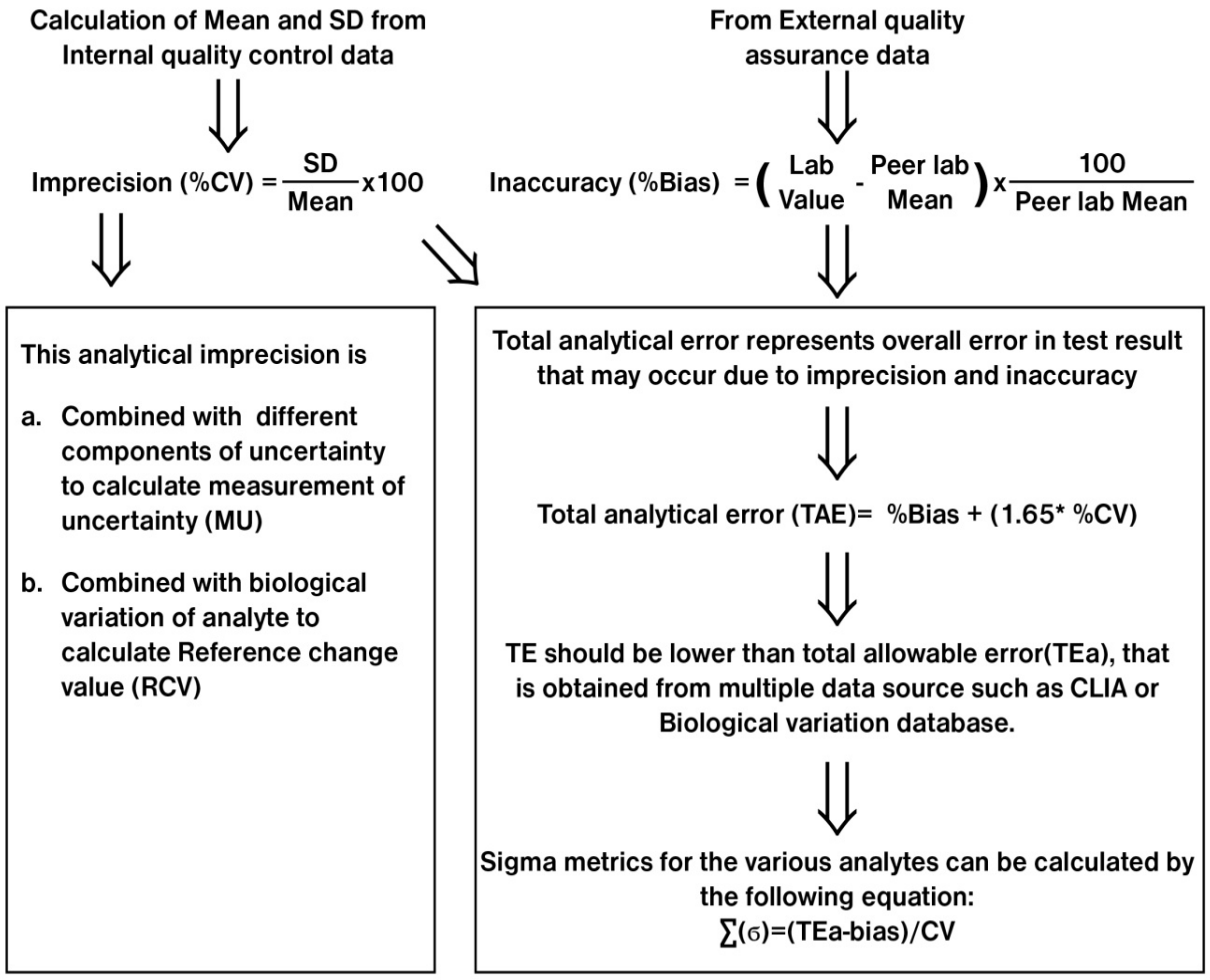
The target SDI is 0.0, which indicates that the labs mean is identical to the consensus group mean. A positive or negative deviation from this target statistic may indicate a bias compared to the consensus group mean.

DIAGNOSTIC POWER OF A TEST

Any user of the laboratory report wants to know the probability of disease given a positive or negative test result. There is no such ideal test which can achieve a perfect discrimination for non-diseased and diseased individuals.

Diagnostic accuracy of a test is measured by calculating the tests' sensitivity, specificity, and predictive values (Figure 4 and 5); these can be

Figure 3 Use of internal and external quality control data



further utilized to construct a Receiver Operating Characteristics (ROC) curve.

Limit of Detection (LoD), and Limit of Quantitation (LoQ) are terms used to describe the smallest concentration of a measurand that can be reliably measured by an analytical procedure. (Figure 5) LoD is the lowest analyte concentration at which detection is feasible. LoQ is the lowest concentration at or above the concentration of LoD and this concentration must be sufficient to produce analytical signals that meet predefined targets for bias, imprecision and total error. LoD is important for tests used to discriminate between the presence and absence of an analyte (e.g. drugs, troponin-I, human chorionic gonadotrophin). Likewise LoQ is important to reliably measure low levels of analyte (e.g. TSH, CRP) for clinical diagnosis and management.

Sensitivity and specificity are not absolute. They are affected by the prevalence of disease and may vary among different populations. Each laboratory test has its defined sensitivity and specificity by the manufacturer and it should be

taken into the clinical consideration for appropriate application of the test.

If a test has high sensitivity, it would not miss a disease, but will also yield false positive results. If a test has high specificity, it will find patients who do not have disease but there will be people who have disease and will be tested negative. This is more dangerous if the investigations are related to infectious diseases. The threshold for a given test is determined by examining the ROC curve, where the sensitivity is plotted as the function of the 1-Specificity for different cut off points. (Figure 6) The area under the ROC curve reflects the diagnostic ability of a test to differentiate people with and without disease of interest.

For example, if the area under the ROC curve is 96%, then there is a 96% chance that a randomly selected diseased person would have a more abnormal result than a randomly selected

Figure 4 **Formulas to calculate diagnostic power of a test**

$$\textbf{Sensitivity} = \frac{\text{True positive}}{\text{True positive + False negative}}$$

$$\textbf{Specificity} = \frac{\text{True negative}}{\text{False positive+ True negative}}$$

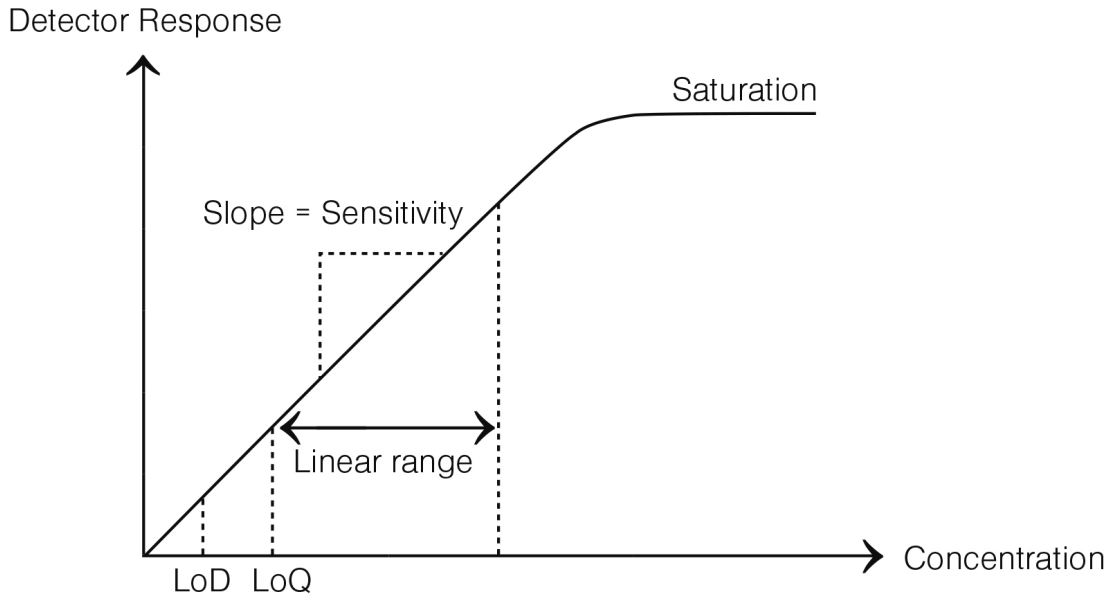
$$\textbf{Positive predictive value} = \frac{\text{True positive}}{\text{True positive+ False positive}}$$

$$\textbf{Negative Predictive value} = \frac{\text{True negative}}{\text{True negative+ False negative}}$$

$$\textbf{Positive Likelihood ratio} = \frac{\text{Sensitivity}}{\text{1-Specificity}}$$

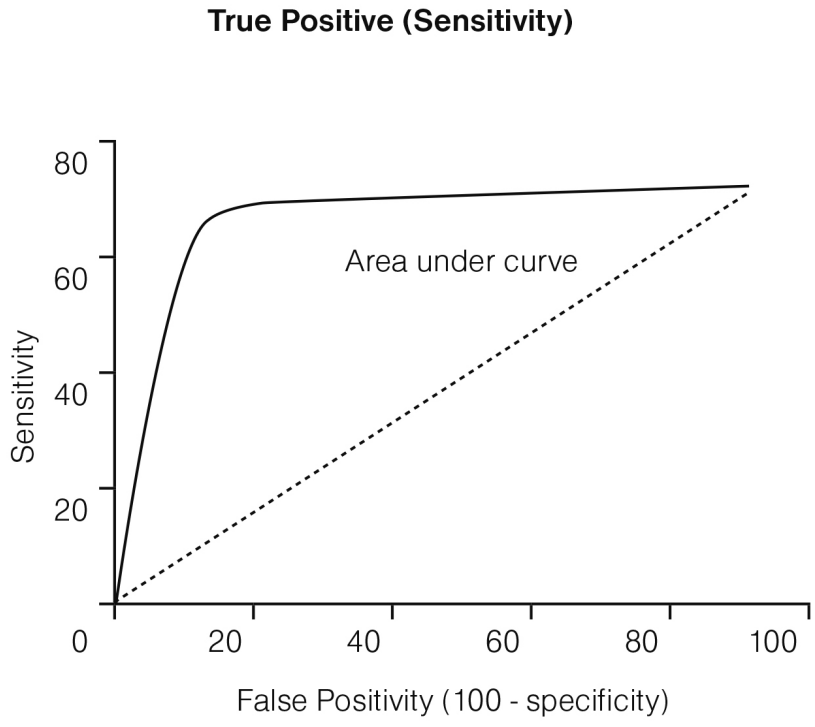
$$\textbf{Negative Likelihood ratio} = \frac{\text{1-Sensitivity}}{\text{Specificity}}$$

Figure 5 **The slope of the concentration versus detector response signifies the sensitivity of the test**



LOD- Limit of detection, LOQ- Limit of quantitation.

Figure 6 **ROC curve**



non-diseased person. The ROC curve also allows comparing the curves (diagnostic accuracy) generated from two or more tests.

Clinicians are more interested to know the predictive value of a test. The predictive value denotes the overall performance of a diagnostic laboratory test in terms of its ability to accurately distinguish the presence of a disease state with a positive test result from the absence of a disease state with a negative test result. (Figure 4). The negative predictive value can be regarded as a reassurance number - when it is very high, the patient can be assured that they don't have disease.

To calculate the predictive values, the 2x2 table is constructed. Predictive values are affected by outcome prevalence. The lower the disease prevalence, lower will be the positive predictive value and this will raise the negative predictive value. Thus, positive predictive value, even for a good test with a high sensitivity, can be poor when there are few persons with the disease. We can also calculate the predictive value using Bayes' Theorem which describes the probability of occurrence of an event related to any condition. [4]

For a laboratory screening tests, particularly where the results of the individual tests are highly variable, a statistical entity known as Multiple of the Median (MoM) is used to report the results. MoM is helpful to estimate the risk for pregnancy complication such as Down's syndrome, neural tube defect, preeclampsia in various weeks of gestation.

Example- Alpha feto protein (AFP) testing is used to screen for a neural tube defect during the second trimester of pregnancy. Because AFP concentrations normally increase during pregnancy, MoM is used to normalize the test result. The MoM is a measure of how far an individual test result deviates from the median value of a large set of AFP results obtained from unaffected pregnancies. For example, if the median AFP result at 16 weeks of gestation is 20 ng/mL and a pregnant woman's AFP result at that same gestational age is 60 ng/mL, then her AFP MoM is equal to 60 divided by 20 (60/20) or 3.0. In other words, her AFP result is 3 times higher than normal.

Calculation for MoM is done by dividing the patient result of particular biomarker by the median result of same biomarker determined by the laboratory. The Mom cut off for each parameter varies by laboratory as it depends on the population characteristics and medical history as well as the analyzer used for making the measurements.

METHOD VERIFICATION

All the invitro diagnostic instruments and reagents that are available must be documented and approved by an official agency. In Europe, the documentation must get a CE mark, and in the United States, an approval procedure by FDA is mandated.

Validation of the products is done at the manufacturer's level to show that the device/reagent is fit for its purpose. This includes measurement of trueness and precision, linearity, chemical interferences, carryover, and risk appraisal. [5] Clinical laboratories usually limit the verification process to compare claims regarding trueness and precision. The other verification criteria may be regarded as inherent to the method/instrument however it depends on the accreditation bodies.

To verify the precision, at least 5 observations during 5 days, in a patients sample or a reference material, are suggested. [6] When the imprecision is obtained from repeated measurements of the same sample and unchanged conditions, it is called the repeatability or within-series variation. If conditions change between estimating the imprecision, for example, from one day to

another or after recalibration of the measurement procedure, the imprecision is characterized as between-series imprecision. Using both these imprecision, the combined or intralaboratory imprecision can be obtained. Statistically, an ANOVA test can also be used to estimate the within- and between-series variation and provides a method to estimate the within-laboratory variation.

To verify bias, laboratories compare a new measurement procedure with previous ones by splitting samples into aliquots. At least 20 numbers of samples in the entire measuring interval is chosen and measured by both methods. [6] Before the statistical evaluation is performed, the scatter plot and difference plots should be carefully studied to identify outliers and are deleted. Statistically, the significance of the difference between the methods is evaluated by the Student t test. This data is used for various more advanced calculations, for example, the regression function, that is, the slope and intercept, and the correlation coefficient. This is discussed further in the method comparison section below.

METHOD COMPARISON

It is mandatory to evaluate analytical methods in the laboratory before their use for patient examinations. In addition to determining experiments for measuring accuracy and precision, it is also necessary to compare the new method to be introduced and other methods in use.

Method comparison involves testing patient samples during a number of different analytical runs by both the new and current methods. In most of the cases, comparison method is the existing method in one's own laboratory or a reference laboratory.

The comparison aims to estimate the constant and proportional differences between the two methods. Various statistical approaches can be used for method comparison procedures. Pearson's correlation coefficient is often used for such comparisons but does not provide appropriate conclusions. The correlation describes the linear relationship between two data sets, but not their agreement, and does not reveal whether there is a constant or proportional difference between the two data sets.

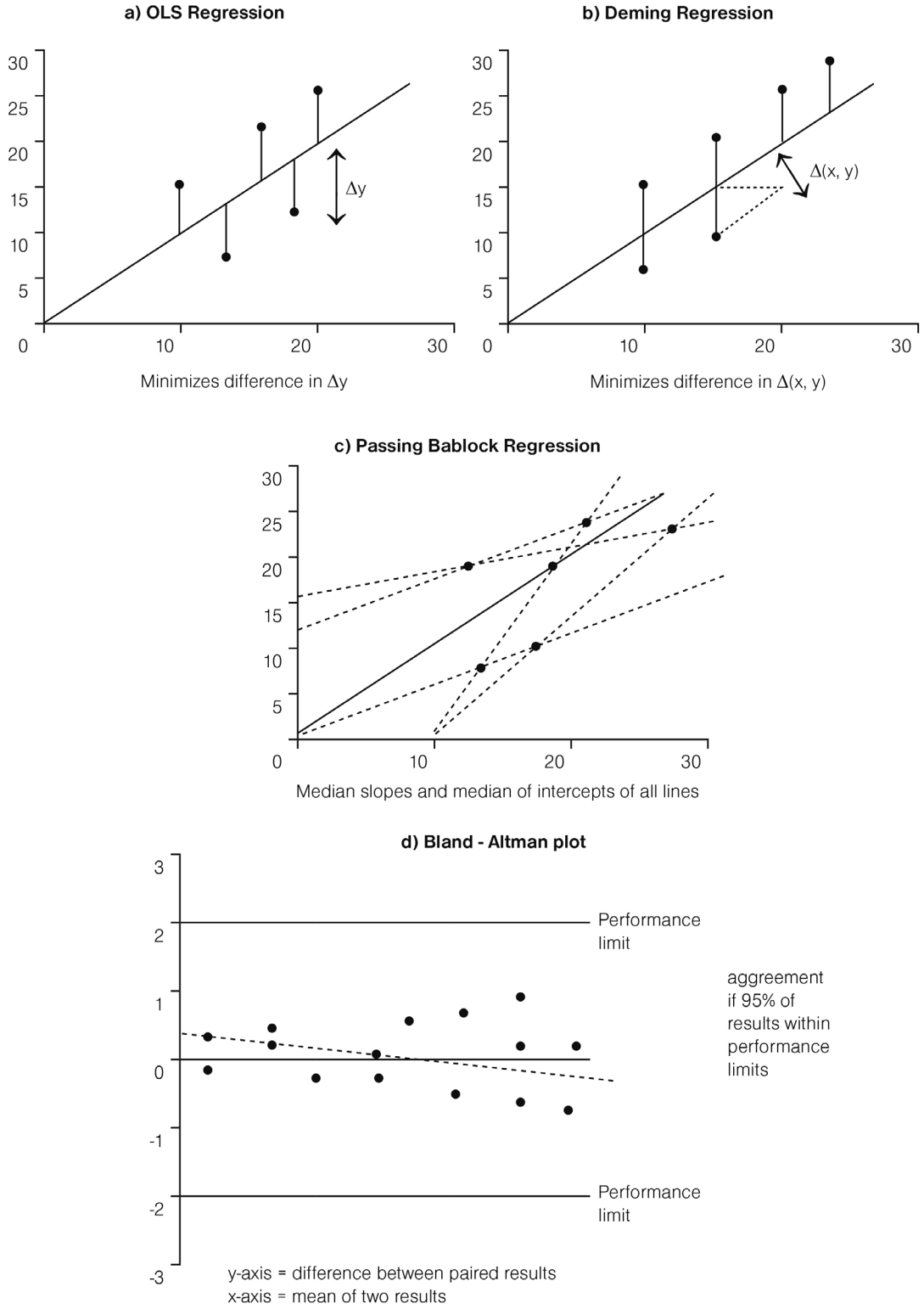
There are various ways to construct the function that binds two variables. To evaluate the equivalences between two methods, a regression function is used. A straight line can adequately describe the relationship between the two variables.

For this purpose, at least 40 patient samples should be analyzed by both methods with at least 2 reagent lots on each analyzer. [7] The analytical concentration should span the entire analytical range. The results are plotted on the Y-axis (dependent) and the reference method (existing) on the X-axis (independent). A linear regression line is inserted through the data points and the slope and Y intercept are calculated. (Figure 7a) There are a number of spreadsheets available that can automatically calculate and plot regression graphs which can be used by the laboratory. [8, 9] The best fit line is defined by the equation; $Y=mx + b$, where m is the slope and b is the Y intercept. A perfect correlation will have all points lying on a line at a 45^{o} angle to the X-axis.

This line will have a Y-intercept of zero and slope of 1. The correlation coefficient (R^2) will be 1.00 and the standard error will be 0.

The common model of this simple linear regression is easy but often may not be suitable for our daily evaluations. The linear regression assumes that the variable x is error-free and that the error of the test method, variable y, is distributed normally and is constant throughout the range of concentrations studied. (Figure 7a) We rarely meet these assumptions in practice.

Figure 7 Illustration of various regression and method comparison models



Thus, other statistical methods for comparing methods have been developed, such as the Passing-Bablok regression, Deming regression, Mountain plot, Bland and Altman plot. (Figure 7b-7d)

Deming regression does not assume that the reference method is free from error and it is the best approach to use when two methods are expected to be identical and the data is normally distributed without outliers. Passing-Bablok regression is used for nonparametric data and performs better when outliers are present. However, Passing-Bablok is computationally intensive and unreliable for small sample sizes.

VARIABILITY IN REPORTS

Serial measurements of laboratory parameter are often required to monitor patient's health. However, repeated laboratory measurements are seldom identical. The change in laboratory result may be due to biological variation, analytical imprecision or a change in patients' health condition. The minimum change required to conclude that two serial measurements are likely different is termed as the reference change value (RCV). A good clinical laboratory should have sufficient data to calculate RCV which are based on the estimates of biological variation (BV) data and analytical variation (AV) data. The BV data are mostly taken from the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) BV database, which delivers real time BV data for numerous analytes. [10] This database is based on results from systemic reviews and published studies by the BV data critical appraisal checklist. [11]

When the pre-analytical conditions are unvarying, the RCV formula becomes:

$$RCV = \sqrt{2} \times Z \times \sqrt{(CVA^2 + CVI^2)}$$

Where, Z indicates the number of standard deviations appropriate to the desired probability, 1.96 for $P < 0.05$; CVA, analytical imprecision; and CVI, within subject biological variation. The CVA of each test is provided by imprecision testing in laboratory.

Acceptable CV or analytical precision needs to be defined for each analyte based on medical significance. Generally, the precision should be equal to or less than one half of the within subject biological variation.

Therefore, analytes with larger biological variation do not require as much analytical accuracy as analytes with small biological variation. For example, BV of fasting triglyceride is 20%; therefore, analytical variation can be as high as 10% without significantly affecting medical decision making.

REFERENCE INTERVAL

When developing reference intervals (RI), clinical laboratories must consider what data sources and statistical methods to use. RI for the same measurements and instruments may differ between laboratories because of the differences in:

a. Operating conditions
b. Criteria for selection of healthy subjects
c. Patient populations
d. Geographical areas in relation to temperature, altitude, barometric pressure and humidity
e. Subject preparation and sample collection

The RI is defined as the interval corresponding to the central 95% of values of a reference population, including the two boundary limits: upper reference limit (+ 2SD) and lower reference limit (-2SD). (Figure 1)

It is recommended that medical laboratories determine their own RIs to cover the variability of their local populations and their specific

analytic methods and devices. For the process of RI determination, the Clinical Laboratory Standards Institute (CLSI) recommends "direct" approach, where well defined reference subjects are selected with pre-defined criteria and the measurements are done afterwards. Direct method is hard to apply for every laboratory in routine practice for it demands much time and money. The alternative approach is the "indirect" method where test results of patients that were ordered for screening, diagnosis or follow-up purposes are derived from laboratory information system (LIS) and used to determine the RIs. This method is faster and cheaper. Besides, the results obtained by the indirect method take into account the analytical and biological variability of the analyzed parameter. Recently, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Committee on Reference Intervals and Decision Limits encourages the use of indirect methods to establish and verify reference intervals.

Both parametric and non-parametric approaches may be taken when analyzing reference range data. The parametric approach involves calculating the mean and standard deviation to determine the range of values that fall within the 95% confidence interval. The non parametric approach involves establishing the values falling at the 2.5 and 97.5 percentiles of the population as the lower and upper reference limits. Outliers can have substantial effect on the calculation of reference ranges by this method and should be removed. Mathematically, outliers are results that differ from the mean by more than 3SD or differ from other results by more than 30%.

Consensus RI for some analytes is determined by medical experts based on the result of clinical outcome studies. Whenever, the consensus RI is available, clinical laboratories should report these values instead of determining their own RI. Example of consensus groups: American Diabetes association, American Heart Association, IFCC etc.

For an FDA approved test method, the clinical laboratories can adopt the manufacturers stated RI. However it should be verified in healthy cohort of samples. Ideally, 40 healthy samples (20 men and 20 women) should be tested and if 95% of the results fall within the published reference range, it can be accepted for use.

CONCLUSION

In this article the basic laboratory statistics is explained in its simplest form. This offers guidance to understand and employ basic statistical controls and methods required by the clinical laboratory.

However, the authors suggest to refer other sources for step-by- step guidance to the quality control, method development, validation/ verification and comparison of test methods.

REFERENCES

1. https://www.aacc.org/cln/articles/2013/september/total-analytic-error Accessed on 19/03/2023

2. Westgard JO, Westgard SA. Total analytic error. From concept to application. 2013.

3. White GH. Basics of estimating measurement uncertainty. The Clinical Biochemist Reviews. 2008 Aug; 29 (Suppl 1):S53.

4. Kallner A. Bayes' theorem, the ROC diagram and reference values: Definition and use in clinical diagnosis. Biochemia medica. 2018 Feb 15;28(1):16-25.

5. Westgard JO. Basic method validation, Westgard QC. Inc., Madison, WI. 2008:168-74.

6. Clinical and Laboratory Standards Institute, User Verification of Precision and Estimation of Bias: Approved Guideline, third ed., CLSI, Wayne, Pennsylvania, 2014. CLSI Document EP15-A3.

7. Clinical and Laboratory Standards Institute, Measurement Procedure Comparison and Bias Estimation Using Patient Samples: Approved Guideline, third ed., Clinical and Laboratory Standards Institute, Wayne, PA, 2013. CLSI document EP09-A3.

8. https://www.westgard.com/. Accessed on 19 March 2023.

9. https://www.acb.org.uk/.Accessed on 19 March 2023.

10. Aarsand AK, Fernandez-Calle P, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, Jonker N, Minchinela J, Simon M, Braga F, Perich C, Boned B, Roraas T, Marques-Garcia F, Carobene A, Aslan B, Barlett WA, Sandberg S. The EFLM Biological Variation Database. https://biologicalvariation.eu/.

11. Aarsand AK, Røraas T, Fernandez-Calle P, Ricos C, Díaz-Garzón J, Jonker N, Perich C, González-Lao E, Carobene A, Minchinela J, Coşkun A. The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation. Clinical Chemistry. 2018 Mar 1;64(3):501-14.