Chapter 5

Systematic Error Detection in Laboratory Medicine

Amir Momeni-Boroujeni and Matthew R. Pincus

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72311

Abstract

Measurements in laboratory medicine have a degree of uncertainty; this uncertainty is often called "error" and refers to imprecisions and inaccuracies in measurement. This measurement error refers to the difference between the true value of the measured sample and the measured value. One of the types of error is systematic error, also called bias, because these errors errors are reproducible and skew the results consistently in the same direction. A common approach to identify systematic error is to use control samples with a method comparison approach. An alternative is use of statistical methods that analyze actual patient values either as an "Average of Normals" or a "Moving Patient Averages." Fundamental questions should be decided before a quality control method is used: how are weights assigned to the results? Is preference given to more recent samples or to the older samples? How sensitive should the model be? In this chapter, we will expand the fundamental notion of systematic error and explain why it is difficult to identify and measure and current statistical methods that are used to detect systematic error or bias.

Keywords: bias, systematic error, measurement uncertainty, bias detection, method comparison, patient average methods

1. Introduction

The role of clinical laboratory is to measure and test patient samples. These measurements are a central part of modern clinical management; they are used by clinicians to diagnose disease states, to guide treatment course and to determine prognosis. The modern clinical laboratory uses a plethora of instruments to quantify and measure different analytes and reports results that are used by clinicians. The most important metrics that a test must possess to be used in clinical laboratory are technical accuracy and precision [1].

A test is technically accurate if it produces valid information. A precise test will produce similar results when the test is repeated multiple times. Accuracy (or rather trueness) is a measure of the

proximity of the test results to the true value. Precision measures reliability and reproducibility. These metrics are complementary and a good clinical test needs to be both accurate and precise [2]. Some have suggested that trueness should be used to refer to the agreement of the measurement to the true value and accuracy to encompass both trueness and precision.

Accuracy and precision are related to a concept called measurement error: every measurement is associated with a degree of error or uncertainty. The goal in laboratory medicine is to minimize the measurement error so that it does not adversely affect the clinical decision-making process. Measurement error can never be truly nullified, but it can be decreased to a scale that is acceptable by clinicians, laboratory directors and laboratory regulatory agencies [2, 3].

Measurement errors can be random, i.e. they can be unpredictable. All measurements have random error. Random errors are due to unpredictable variations in sample, instrument, measurement process or analysis and it can be said to follow a Gaussian distribution, i.e. random error follows randomness and chance and thus laws of probability apply to random error. As the instruments get more precise the Gaussian distribution of the random error gets narrower and the random error decreases. At the same time, if we repeat an experiment or test multiple times we can average out random error from our measurements. i.e. the mean of multiple repeated measurements gets closer to the true value as the number of repeats increases. This forms the basis of reporting confidence intervals for measurements [2, 4].

Bias or systematic error is a form of measurement error that skews the results to one side. Repeating the measurements cannot eliminate bias. In other words, bias is a non-zero error which will consistently affect the results and can show a problem with the measurement process often requiring corrective action. The corrective action can be in form of calibration by introducing a correction factor or by changing components of measurement. Systematic error can be short-term or long-term, with very short-term systematic error often manifesting as random error.

Systematic error and random error have a cumulative effect on the measurement results (**Figure 1**). Thus, measurement error is often considered as total error with both bias and random error contributing. Laboratories often have limits for total error, bias and random

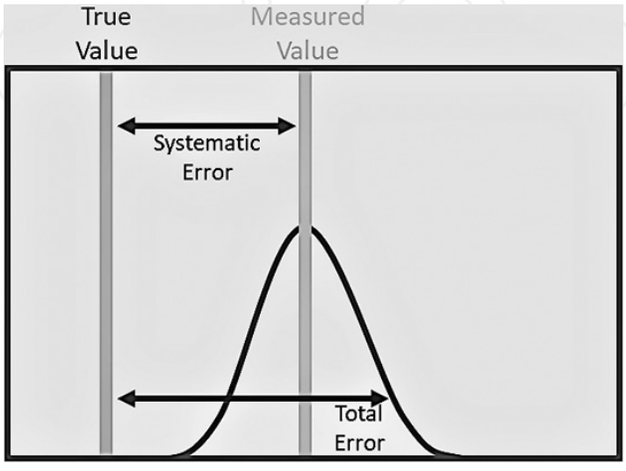

Figure 1. This figure depicts the cumulative effect of systematic error and random error. The X-axis represents the value determined and the Y-axis plots the frequency of occurrence of each value.

error. All tests need to be checked continuously for presence of error and identifying systematic error is part of the function of a clinical laboratory. The measurement error can be regarded as a noise that can obscure the signal or true test value. In the presence of noise, drawing conclusions from the signal that may change the true value in a clinically significant manner risks jeopardizing the patient's health. As a result, the lab should strive to identify noise, minimize it or reduce its impact on patient outcomes. In this regard, systematic error is especially dangerous since it will skew the test results in a manner that cannot be corrected by repeat measurements. Unfortunately, systematic error can be very difficult to identify and/or quantify. In this chapter we focus on approaches for identification of systematic error using within-laboratory comparisons [5, 6].

2. Systematic error detection using quality control experiments

Simply stated, the aim of quality control experiments is to determine the performance of the laboratory tests with measuring of known samples or references, that is, samples in which the true value of the analyte being tested is known. These methods are mainly set up to detect random error and check instrument precision. However, the same results can be used to detect bias and systematic error [7].

The laboratories can use certified reference materials to measure and identify systematic error. If the reference sample is measured with each analytical run, you would expect the results of the reference sample measurements to show a random distribution around the true value, yet if the results are consistently lower or higher than the reference value then you would suspect that a bias exists [2, 8].

For systematic error measurement, a method comparison method is needed to identify systematic error. Any systematic error found needs to be corrected using a recovery experiment and calibration.

2.1. Levey-Jennings plots

The first step in identification of systematic error is to visually inspect the quality control process. Levey-Jennings plot shows the fluctuation of reference sample measurements around the mean against time. The chart's reference lines include control limits, 2 standard deviation lines, 1 standard deviation lines and the mean reference line.

The mean, standard deviation and the control limits are calculated by a replication study where the certified reference material is repeatedly measured. The repeated measurements allow for calculation of mean and standard deviation of the control sample levels. The trial limits are mean ± 3 standard deviations. The next step is to eliminate the replication study results that are beyond the 3 standard deviations. Then the mean and standard deviation are recalculated and the trial limits are again set. Again, results beyond the trial limits are excluded. The process continues until all the remaining results are within the trial limits. These final trial limits, mean and standard deviation are set as the reference measures for that reference sample.

The number of replication studies to perform can be calculated based on the number of acceptable failures. The sample size calculation is based on set levels of confidence and reliability. Confidence (accuracy) is the difference between 1 and type I error rate. Reliability is the degree of precision. For a failure rate of 0 (i.e. we are not allowing any incorrect results), the equation can be stated as:

$$n = \frac{\ln(1 - confidence)}{\ln(reliability)} \tag{1}$$

The confidence level is often set at 0.95 and reliability at 0.90 or 0.80. If we allow failure events, then the calculation of the sample size is based on the following equation:

$$1 - Confidence = \sum_{i=1}^{f} \binom{n}{i} (1 - Reliability)^i Reliability^{n-i} \tag{2}$$

where f is the failure rate and n is the sample size.

In a Levey-Jennings plot the X-axis represents time and Y-axis represents the measured value. Reference lines are drawn parallel to the X-axis corresponding to mean, mean ±1 standard deviations, mean ±2 standard deviations, and mean ±3 standard deviations. The next step is to plot measured values of the reference material for each run on the plot (**Figure 2**).

2.2. Westgard rules

Westgard rules are a set of guidelines set by Dr. James Westgard for identification of random and systematic error in laboratory quality control experiments. They are based on repeated measurements of at least two reference samples with each analytical run. Some of the Westgard rules are

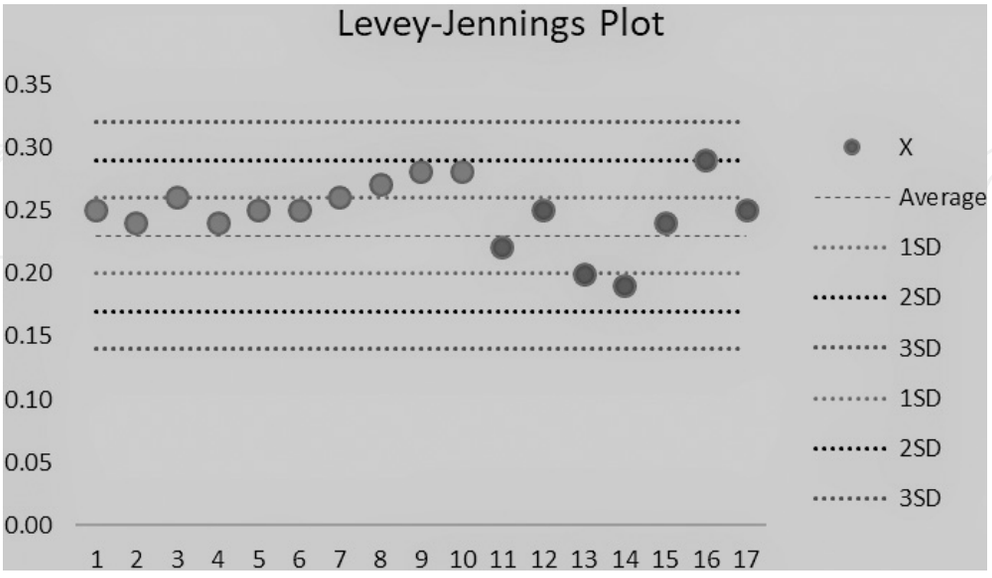

Figure 2. An example of a Levey-Jennings plot. X-axis plots the time of measurement (e.g. day) and the Y-axis plot the measurement value for that unit of time. The lines denoting the mean value and 1, 2 and 3 standard deviations from the mean are explained in the figure.

concerned with identification of random error and within runs error detection [2, 7]. Other Westgard rules are focused on identification of systematic error and between runs error detection. In this chapter we will focus on the latter rules.

- 2_{2S} rule: The QC results are considered to have failed and a bias is present if two consecutive control values fall between the 3 standard deviations and 2 standard deviation limits on the same side of the means reference line.
- 4_{1S} rule: The QC results is considered to have failed and a bias is present if four consecutive control values fall on the same side of the mean reference line and are at least one standard deviation away from the mean.
- 10_{x} rule: The QC results are considered to have failed and a bias is present if 10 consecutive control values fall on the same side of the mean reference line.

These rules are shown in **Figure 3**.

2.3. Method comparison

Method comparison is used for initial assay validation as well as for studying accuracy of a test. The aim of method comparison is to establish whether the assay measures what it is supposed to measure and how accurately it measures it. The findings of method comparison also allow for correction of the results if a bias is found (i.e. calibration). The principal for method comparison is that a gold standard or a standard reference material exists where in the amount of analyte in the sample is exactly known (or known with a high degree of accuracy). We can use this reference standard as a comparator against the performance of our assay and determine the degree of bias that exists in our measurements. This essentially means that we are measuring the relative performance of our assay against the reference standard.

Ideally, identification of a bias should lead to a search for the source of the bias and systematic error, and attempts should be made to rectify the cause of the observed bias. However, there are instances in which no fault or solvable problem is identified; in these instances, if the assay has enough precision and stability as well as clinical merit then we can use the findings of method comparison to adjust for the observed bias.

Bias can take two general forms: constant bias and proportional bias. The constant bias is a difference between the observed measurement and the expected measurement that is constant throughout the range of the observations. Constant bias ($\beta 0$) is represented in regression statistics

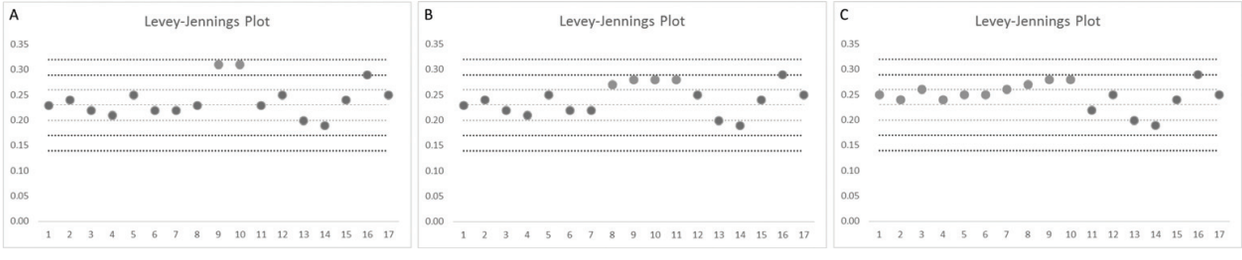

Figure 3. Examples of systematic error in Levey-Jennings plot: A. An example of 2-2S rule, B. An example of 4-1S rule, C. An example of 10x rule.

as intercept. Proportional bias ($\beta 1$), on the other hand, is proportional to the observed value of the measurement and varies across the range of measurements. Proportional bias is represented in regression statistics as the slope of the regression line. If the expected value of measurement is Y_i for each sample i, and the observed value of measurement for sample i is X_i, then we can form a linear regression between the expected values and observed values:

$$Y_i = \beta 0 + \beta 1 \, X_i + \varepsilon_i \tag{3}$$

where ε_i is the random error of the expected observations under the Youden assumption which states that the random error of observed values is smaller than the random error for expected values.

The regression formula is the representation of the best regression line that shows the relationship of the observed value to the expected value. **Figure 4** shows the regression lines for different constant and proportional bias levels.

If no bias exists then $Y_i = X_i$.

The simple linear regression formula allows us to calculate the constant and proportional bias using a simple unweighted ordinary least squares estimator. In ordinary least squares (OLS) models, different candidate values for the parameter vector $\beta 1$ are tested to create regression lines. Then for each i-th observation the residual for that observation is calculated by measuring the vertical distance between the data point (Y_i, X_i) and the regression line formed using the candidate value. The sum of squared residuals (SSR) is determined as a measure of the overall model fit. The candidate value that minimizes the sum of squared residuals is considered as the OLS estimator for the slope. For simple method comparison studies where only two comparators are present the model can be simplified as:

$$\beta 1 = \frac{\sum X_i Y_i - \frac{1}{n} \sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n} (\sum X_i)^2} = \frac{Covariance(X, Y)}{Variance(X)} \tag{4}$$

The constant bias can be calculated by subtracting the mean expected value from mean observed value weighted by proportional bias:

$$\beta 0 = \overline{Y} - \beta 1 \, \overline{X} \tag{5}$$

Constant and proportional bias usually has different root causes. Constant bias often stems from insufficient blank sample correction and is fairly easy to address and rectify. Proportional

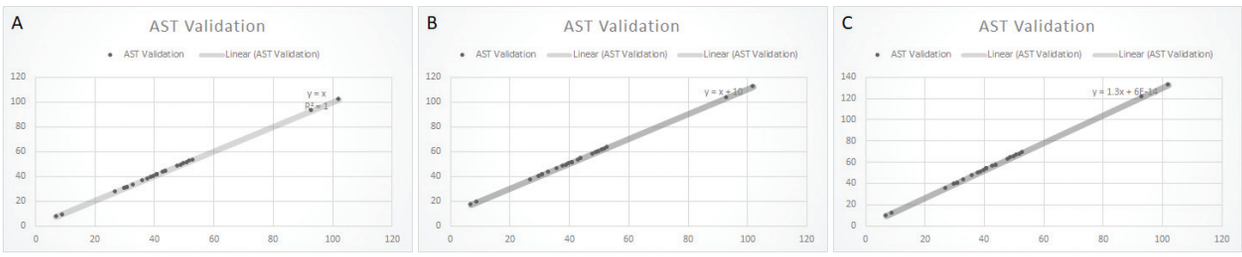

Figure 4. A. When no systematic error exists. B. Shows constant bias. C. Shows a proportional bias.

problems can sometimes be caused by the difference in the composition of calibrator samples and the standard samples or biologic test matrices. The matrix of the reference standard is usually near the actual matrix of the patient samples and thus may contain confounders which may adversely affect the measurement. Yet calibrators often do not have a biologic matrix. If the source of the proportional bias is due to calibration problems, then a recalibration can rectify the problem.

The problem with the Youden assumption is that it considers our observations to have no random disruptions, an assumption which is false as we know every measurement is associated with a degree of uncertainty and imprecision. Alternatively, we can use Deming's regression where the random error for both expected and observed values is factored into the calculation of the proportional and constant bias. In Deming's regression a ratio of the variances of the random error of observed and expected values is calculated:

$$\delta = \frac{{\sigma_\varepsilon}^2}{{\sigma_\eta}^2} \tag{6}$$

where ${\sigma_\varepsilon}^2$ is the variance of the expected values random error and ${\sigma_\eta}^2$ is the variance of the observed values random error. Using this ratio, the OLS estimator for the proportional bias can be given by:

$$\beta 1 = \frac{(Var(Y) - \delta Var(X)) + \sqrt{(Var(Y) - \delta Var(X))^2 + 4\delta Covar(X,Y)^2}}{2Covar(X,Y)} \tag{7}$$

This regression formula is also known as the maximum likelihood estimator [9].

If a linear relation between errors and measurements exists (or is assumed) then an alternative method for error detection is to create Bland-Altman plots. In these plots, the average of the paired values for expected and observed values is plotted on the x-axis and the difference of each pair is plotted on the y-axis. In this method the average difference of the values is called bias and the standard deviation of the differences is also calculated to determine the limits of agreement which constitutes Mean difference ±1.96SD.

The Bland-Altman approach allows for a visual inspection of the proportional bias. However, by dividing the limits of agreement by the mean value of the expected values we can obtain a metric called percentage error. The acceptable percentage error levels for different analytes have been determined and are standardized. In cases where the percentage error exceeds the acceptable levels, corrective action is needed for the detected bias [10].

2.4. R statistics

One of the important statistics for simple linear regression is calculation of the Pearson's r coefficient. This coefficient shows how well the compared results change together and can have values of between minus 1 and 1. This coefficient can be calculated by dividing the covariance of the two variables to the product of their standard deviations:

$$r = \frac{Covar(X, Y)}{\sigma_X \sigma_y} \tag{8}$$

The closer the r coefficient gets to 1, the greater the linear relationship is between the two variables. Some interpret the r coefficient as a measure of correlation with r coefficients more than 0.8 showing correlation. However, in laboratory medicine a correlation of 0.8 actually signifies a great degree of bias. In fact, laboratories should aim for a perfect degree of linearity (r > 0.99) to ensure that systematic error is minimized. Attaining a Pearson's r coefficient of <0.975 signals the presence of systematic error and should prompt the lab to conduct further investigation (using t-test and f-test) to determine the source of this error.

The degree of agreement or the coefficient of determination (R^2). This coefficient is calculated from the ratio of explained variance to the total variance of Y:

$$R^2 = \frac{\sum \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum \left(Y_i - \overline{Y}\right)^2} \tag{9}$$

where $\widehat{Y}_i$ is the calculated value of Y based on the regression for the i-th observation and Y_i is the actual value of Y for i-th observation.

Alternatively, the coefficient of determination can be simply calculated by squaring the Pearson's r coefficient. While the Pearson's r coefficient shows the presence of linearity, the coefficient of determination helps us to determine how well the regression line fits the actual data points. In assessment of a method comparison evaluating this coefficient is necessary as it shows fit of the model: The closer the coefficient gets to 1, the better the regression line fits actual data points. However, it must be noted that even at numbers very close to 1 significant bias may exist. For example, a 5% bias will only result in a R squared score of 0.99 and a 10% bias will result in a R squared score of 0.96. For laboratory medicine purposes we should aim for a R squared score of more than 0.99.

2.5. T-test and F-test

In cases where there is a suspicion of significant bias (as determined by Pearson's r or R squared statistics), then we should determine whether the bias stems from difference in the mean assay concentration or in the variance of the assay. To check for mean we run a paired t-test, and, to check for variance, we run an f-test.

The paired t-test is performed by comparing the means of the observed and expected values; more specifically the mean difference of the values (μ_D) is used for the comparison. The t-statistics can be calculated by:

$$t = \frac{\mu_D}{\sigma_D / \sqrt{n}} \tag{10}$$

where n is the number of data points and σ_D is the standard deviation of the mean difference. To determine the significance of the results (the p-value), the t-statistics should be looked up on a t table corresponding the degree of freedom; the degree of freedom in paired t-tests equals n–1.

A t-test with a significant p-value signifies the presence of a significant bias in the mean of the methods. The next step then would be to determine whether the systematic error represents a constant bias or a proportional bias. This can be done by examining the regression curve or equation. The presence of an intercept signifies a constant bias while presence of a slope other than 1 signifies proportional error. The correction for a constant bias is simple and would require adding the constant to the measurement results. Correction of the proportional bias, however, requires a recovery experiment as described in Section 3.8 below.

The f-test compares the expected variance of the values to the observed variance; while the t-test compares the centroid of the data points (the mean), the f-test deals with the distribution and variance of the data points (the variance). The t-test is more sensitive to differences in the values in the middle of the data range while f-test is more sensitive to differences in the extremes of the data range. A significant f-test would signify random error in the measurement or in other words imprecision. To calculate the f-test the following equation is used (the larger of the two variances will always be the numerator and the smaller one the denominator in this fraction):

$$f = \frac{Var_1}{Var_2} \tag{11}$$

The degree of freedom of the f-test is (n-1, n-1) and the significance threshold can be looked in a f-table corresponding the degree of freedom.

It is important to perform the f-test prior to the t-test; one of the basic assumptions of the t-test is that the standard deviations of the data points are similar between the two groups, i.e. no significant imprecision should exist for t-test results to be valid. In presence of a significant imprecision, the determination of presence of a significant bias should be done using a Cochran variant of the t-test.

In Cochran variant of t-test, standard deviation cannot be pooled between the two groups:

$$t = \frac{\mu_D}{\sqrt{\frac{Var_1 + Var_2}{n}}} \tag{12}$$

The critical value for the t-statistics should also be calculated:

$$Critical\ t = \frac{\frac{t}{n}(Var_1 + Var_2)}{\frac{Var_1 + Var_2}{n}} \tag{13}$$

where t is the t-score corresponding to n-1 degrees of freedom [11].

2.6. Accuracy profile

Accuracy profiling has moved away from treating bias and imprecision as separate entities. In fact, most guidelines (whether based on the total error principles or measurement uncertainty principles) combine bias and imprecision for acceptability criteria. To calculate bias and

imprecision, we need to run a reproducibility study. Reproducibility of quantitative studies is obtained by repeated measurements of a sample in a series and then conducting multiple series of reproducibility studies.

The overall measurement of bias will be the difference between the mean value of the analyte obtained from the repeated measurement and the reference value:

$$Bias = Overall\ mean - Reference\ value \tag{14}$$

Bias and imprecision are used to form the tolerance interval; it is the interval which, with a determined degree of confidence, a specified proportion of results for a sample fall. Tolerance interval can be expressed as:

$$Tolerance\ Interval = reference\ value + bias \pm intermediate\ precision \tag{15}$$

For laboratory medicine, the tolerance interval of analytes needs to be smaller than the acceptability limits. In united states, the acceptability limits are set and governed by the Clinical Laboratory Improvement Amendments of 1988 (CLIA88). These acceptability limits are provided under the following heading: 42 CFR Part 493, Subpart I - Proficiency Testing Programs for Nonwaived Testing (https://www.gpo.gov/fdsys/pkg/CFR-2011-title42-vol5/pdf/CFR-2011-title42-vol5-part493.pdf).

The important factor from intermediate precision that is needed in calculation of tolerance interval is the standard deviation of reproducibility (S_R). The standard deviation of reproducibility can be calculated by the following equation:

$$S_R{}^2 = \frac{1}{n}\left(\frac{Var_{betweenseries}}{p-1} + (n-1)\frac{Var_{withinseries}}{n-p}\right) \tag{16}$$

where n is the number of within-series measurement repeats and p is the number of series of reproducibility measurements.

An advantage of calculating the intermediate precision is that we can use it in combination with within- series repeatability to determine the uncertainty of bias:

$$Uncertainty\ of\ Bias = 1.96\left[\frac{n\left(S_R{}^2 - S_r{}^2\right) + S_r{}^2}{np}\right]^{1/2} \tag{17}$$

$S_r{}^2$ is the within-series repeatability and can be calculated using the following equation:

$$S_r{}^2 = \frac{Var_{withinseries}}{p(n-1)} \tag{18}$$

Uncertainty of bias is essentially 1.96 times the standard deviation of bias which corresponds to a 95% confidence interval for bias determination.

The between-series reproducibility is calculated using the following equation:

$$S_L{}^2 = \frac{1}{n}\left(\frac{Var_{betweenseries}}{p-1} - S_r{}^2\right) \tag{19}$$

The between-series reproducibility is used in calculation of the Mee factor (K_s). Mee factor is the other component of intermediate precision. Since the calculation of the Mee factor is complicated we have broken it down into a series of equations. The first step is to calculate the H ratio:

$$H = \frac{S_L{}^2}{S_r{}^2} \tag{20}$$

The next step is to calculate the G^2:

$$G^2 = \frac{H+1}{nH+1} \tag{21}$$

Which in turn is used to calculate C:

$$C = \left(1 + \frac{1}{npG^2}\right)^{1/2} \tag{22}$$

The final step is to multiply C by the t-score associated with the degree of freedom (dof):

$$Degree\ of\ Freedom = \frac{(H+1)^2}{\frac{\left(H+\frac{1}{n}\right)^2}{p-1} + \frac{1-\frac{1}{n}}{np}} \tag{23}$$

And:

$$K_s = C \times t_{dof} \tag{24}$$

By calculating the Mee factor and the standard deviation of reproducibility we can now obtain the intermediate precision:

$$Intermediate\ precision = K_s \times S_R \tag{25}$$

Thus, we can rewrite the tolerance interval as [12]:

$$Tolerance\ Interval = reference\ value + bias \pm (K_s \times S_R) \tag{26}$$

2.7. Weighting procedures

The problem with simple linear regression is that is based on a set of assumptions; one of the problematic assumptions is that the standard deviation of the random error is constant throughout the range of measurement. This assumption, however, is often wrong as the standard error of measurement is often much larger near the extremes of measurement range (near the limit of detection and the highest range of linearity). The solution in laboratory

medicine can be to run linearity experiments and limit the measurement range based on the linearity results. Despite this the effect of random variation on the regression line remains. To rectify this, a solution is to employ a weighting procedure.

The simplest weighting procedure is to use the standard deviation of variation for each data point of the method comparison study. This requires that the method comparison study is repeated multiple times (20-30 times). This allows us to calculate the standard deviation of measurement for each point (S_i). The weighting coefficient will then be the inverse of this standard deviation:

$$w_i = \frac{1}{S_i} \tag{27}$$

This weight can then be incorporated into the equations of the method comparison. For example, the r coefficient can be recalculated as:

$$r = \frac{\sum w_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\left(\sum w_i \left(X_i - \overline{X}\right)^2 \sum w_i \left(Y_i - \overline{Y}\right)^2\right)^{\frac{1}{2}}} \tag{28}$$

Weighting can often considerably decrease the bias percentage especially at the extremes of measurement compared to non-weighted regression. Weighting by inverse of standard deviation tends to normalize the relative bias at the extremes of measurement while weighting by inverse of variance tends to favor the bias correction for lower ends of measurement (less bias at lower concentrations). The decision for weighting and/or choice of weighting procedure should be based on the assay characteristics and performance requirements [13].

2.8. Recovery percentage

To estimate the proportional bias, a recovery experiment is needed. The recovery experiments are performed by calculating the amount of recovery when adding a known amount of the analyte to the sample: this is done by dividing the measurement sample into two equal aliquots and performing the measurement for both aliquots. To one of the aliquots, a known amount of target analyte is added (aliquot 1). For the other aliquot (aliquot 2) an equal amount of diluent is added and the measurement is repeated. The recovery percentage can then be calculated:

$$Recovery\% = \frac{(Analyte\ amount\ in\ aliquot\ 1) - (Analyte\ amount\ in\ aliquout\ 2)}{Amount\ of\ analyte\ added\ to\ aliquote\ 1} \times 100 \tag{29}$$

The recovery or bias percentage is often used in laboratory medicine to state the proportional bias. Most of the regulatory agencies have set critical values for the recovery percentage for different analytes. The advantage of using recovery percentage is that it normalizes to 100 allowing for easier understanding of the scale of bias present [2].

3. Bias detection without comparators

Up to this point we have discussed bias detection methods that use a reference material or comparator to assess the presence of bias. While this has been the accepted standard for many laboratory regulatory agencies, there are arguments against this approach to bias detection: first of all, the assumption of method comparison studies is that the reference material (control samples) values are true and do not suffer from imprecisions. The measurement uncertainty is considered to be minimal in these samples. Yet, unless these samples vary considerably from the biologic sample matrix, a degree of measurement uncertainty would exist in these samples which lead to inaccurate estimates of bias and imprecision of laboratory instruments and techniques. On the other hand, running repeated control samples with each run and the need for revalidation of the instrument and techniques after each change in the parameters, requires a considerable investment in terms of time, labor and cost.

Alternatively, the systematic error can be determined by using the patient samples. This can be done by either tracking the results of known normal patients (i.e. those expected to have a result within the reference range based on their clinical and physiologic state) or by following the trend of all the results of an analyte over time. Using patient samples has the advantage of including the inherent biologic uncertainty into the calculation of bias.

3.1. Average of normal (AON)

In this approach the comparator for quality control would the average values of the analyte in normal individuals. This requires us to know the population average and standard deviation for that analyte. If we measure the analyte in a normal individual, we would expect the results to approximate the population average. Deviations of the normal results from the expected reference normal can signal the presence of a systematic error.

In AON, the mean value of normal samples is compared to a mean reference value. The mean reference value should be established by the laboratory based on the population it serves; this is best done as part of the initial validation of an assay when a large size sample of normal individuals is tested to establish the reference ranges. This experiment allows us to calculate the population mean, standard deviation and standard error (SD/√N). We expect the Average of Normals from our analytical run to fall within the 95% confidence interval of the population mean.

$$95\%CI = Population\ Mean \pm 1.96\ Standard\ Error \tag{30}$$

With each analytical run, a sample of normal results should be used to calculate the Average of Normals for that analytical run. If the calculate average is beyond the 95% CI of the population then we have detected a systematic error in the analytical run.

In AON method, as the size of the normal sample increases the probability of detecting bias also increases. The size calculations for the AON method are determined by the ratio of the biological variance of the target analyte (CVb) to the variance of the method (CVa) (CVb/CVa)

as well the expected probability of detecting the bias. To help with these calculations, one can utilize the Cembrowski nomogram [14] or, alternatively, the methods used in [15]. It is also possible to perform the AON by performing a two-sample independent t-test.

3.2. Moving patient averages

Unlike the AON method, in moving patient averages, all the results of an assay are included in evaluation of bias. The principle for moving patient averages is that the samples tested in a laboratory follow a repeating pattern. This assumption means that the overall biologic and clinical spectrum of patients and individuals tested in the laboratory is constant throughout the analytical runs. In moving patient averages, we expect the average results of an assay for two overlapping subsets of patient to be constant. In this method, for example, an average is calculated on the first 100 patients, should be similar to the average calculated based on the results of patients number 2 to 101, etc.

The moving average can be calculated using exponentially weighted moving average ($\overline{X}_{M,i}$). It is important to consider that, in moving patient averages the weight $(1-r)$ assigned to previous results average ($\overline{X}_{M,i-1}$) should be greater than the weight (r) assigned to the most recent results ($\overline{X}_i$) (in other words the average of each batch is weighted down by previous averages). This can be stated as:

$$\overline{X}_{M,i} = r\overline{X}_i + (1-r)\overline{X}_{M,i-1} \tag{31}$$

The weight assigned to current values is usually set between 0.05 and 0.25 with recommended value of 0.1.

The comparator in moving patient averages are the control limits. We expect the weighted patient average to fall within the control limits for that test. Any moving patient average outside of this control limit signifies the presence of a bias. The control limit equation is provided below.

$$\textit{Control limits of exponential moving average} = \overline{X}_{M,0} \pm L\sigma\sqrt{\left|\frac{r}{2-r}\left[1-(1-r)^{2i}\right]\right|} \tag{32}$$

where L is a constant set based on the confidence level (for 95% CI, L equals 2), and σ is the standard deviation of the current batch.

The moving patient averages can also be evaluated using the Bull's algorithm. In this approach, the moving average ($\overline{X}_b$) is calculated for subsets of 20 samples with 19 patient values and one value representing the previous moving average. These values are weighted differently (i.e. more weight is assigned to the previous moving average than the 19 new samples).

The general formula for Bull's moving average can be written as:

$$\overline{X}_{b,i} = (2-r)\overline{X}_{b,i-1} + rD \tag{33}$$

where $\overline{X}_{b,i}$ is the current moving average, r is the weight for current values (with possible values of $0 < r \leq 1$, usually set to 1), $\overline{X}_{b,i-1}$ is the previous moving average and D is calculated

from the value of current measurements in the batch. If we assume a value of 1 for r then we can write the bull's algorithm as:

$$\overline{X}_{b,i} = \overline{X}_{b,i-1} + \left(\frac{\sum_{j=1}^{N} \sqrt{X_j - X_{b,i-1}}}{N} \right)^2 \tag{34}$$

where N is the number of results in the batch.

The control limits of Bull's moving average are set as $\overline{X}_{b,0} \pm 3\%\overline{X}_{b,0}$, with $\overline{X}_{b,0}$ being the target value for that analyte.

The advantage of moving averages is that they can filter out outliers' effect thus removing confounding by imprecision.

The moving patient averages algorithms are very powerful for detection of bias: they can routinely identify bias percentages of 1% and more. Most automated hematology analyzers use moving patient averages to check for presence of bias in their assays. However, the patient moving averages algorithms have suffered from implementation problems and are not widely used beyond hematology analyzers [2].

3.3. Time series analysis and forecasting for bias identification

An extension of the moving patient averages is the application of time series analysis and forecasting for bias detection. In time series analysis the previous trends of the analyte results are used to predict (forecast) the trend in future. If the observed analyte results deviate from the forecasted trend, then a measurement error may exist. In the setting of laboratory medicine, we need to be able to detect bias in short time series and distinguish the measurement error from the noise and chaos stemming from biologic variation. Here, we will introduce the concept of using time series analysis for bias detection but we will not explain the methodology in depth as it goes beyond the scope of this chapter.

In forecast models, a series of data points are used to create one or more projection patterns for future trends. This is done using forecasting models such as ARIMA (Autoregressive integrated moving average). These projections are often correct for very short-term predictions (next 1 or 2 data points), but for forecasting further, the noise and chaos cause the prediction accuracy to fall. However, by examining the correlation of predicted and observed values and documenting its changes as we forecast further into the future, we can determine if the observed pattern represents the deterministic chaotic nature of biologic measurement or if it represents a measurement error; for measurement error we expect the correlation coefficient to remain constant with time; however, with chaos, we expect the correlation coefficient to deteriorate over time [16].

There are other approaches using times series analysis that can be helpful in systematic error identification. One of these approaches uses unit root tests such as the Dickey-Fuller test [17]. These tests examine whether a time series is stationary over time, i.e., whether the mean and

variance are constant over time. In contrast, nonstationary time series will have either a varying mean and/or varying variance over time. Using this approach any departure from stationarity can signal either a drift (proportional bias) and/or a shift (constant bias) or even increase in imprecision over time (difference-stationary nonstationarity) [17]. If the Dickey-Fuller test returns a significant p-value then we can say that the series is stationary, and no significant measurement error is present.

Author details

Amir Momeni-Boroujeni* and Matthew R. Pincus

*Address all correspondence to: mrpincus2010@gmail.com

Department of Pathology, State University of New York, Brooklyn, NY, USA

References

[1] McPherson RA, Pincus MR. Henry's Clinical Diagnosis and Management by Laboratory Methods E-Book. Philadelphia, PA, USA: Elsevier Health Sciences; 2017

[2] Momeni A, Pincus MR, Libien J. Introduction to Statistical Methods in Pathology. Switzerland: Springer, Cham

[3] Howanitz PJ. Errors in laboratory medicine: Practical lessons to improve patient safety. Archives of Pathology and Laboratory Medicine. 2005;**129**(10):1252-1261

[4] Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: A practical guide for biologists. Biological Reviews. 2007;**82**(4):591-605

[5] Bonini P, Plebani M, Ceriotti F, Rubboli F. Errors in laboratory medicine. Clinical Chemistry. 2002;**48**(5):691-698

[6] Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017;**355** (6325):584-585

[7] Westgard JO, Westgard SA. Measuring analytical quality. Clinics in Laboratory Medicine. 2017;**37**(1):1-3

[8] Oosterhuis WP, Bayat H, Armbruster D, Coskun A, Freeman KP, Kallner A, Koch D, Mackenzie F, Migliarino G, Orth M, Sandberg S. The Use of Error and Uncertainty Methods in the Medical Laboratory. Clinical Chemistry and Laboratory Medicine (CCLM); 2017. DOI: https://doi.org/10.1515/cclm-2017-0341

[9] Strike PW. Statistical Methods in Laboratory Medicine. Oxford, UK and Waltham, Mass, USA: Butterworth-Heinemann; 2014

[10] Hanneman SK. Design, analysis and interpretation of method-comparison studies. AACN Advanced Critical Care. Oxford, UK and Waltham, Mass, USA. 2008;**19**(2):223

[11] Guidelines for Quality Management in Soil and Plant Laboratories. No. 74. Food & Agriculture Org.; 1998

[12] Mermet JM, Granier G. Potential of accuracy profile for method validation in inductively coupled plasma spectrochemistry. Spectrochimica Acta Part B: Atomic Spectroscopy. 2012;**76**:214-220

[13] Mermet JM. Calibration in atomic spectrometry: A tutorial review dealing with quality criteria, weighting procedures and possible curvatures. Spectrochimica Acta Part B: Atomic Spectroscopy. 2010;**65**(7):509-523

[14] Cembrowski GS et al. Assessment of "average of normals" quality control procedures and guidelines for implementation. American Journal of Clinical Pathology. 1984;**81**(4):492-499

[15] Westgard JO, Smith FA, Mountain PJ, Boss S. Design and assessment of average of normals (AON) patient data algorithms to maximize run lengths for automatic process control. Clinical Chemistry. 1996;**42**(10):1683-1688

[16] Sugihara G, May RM. Nonlinear forecasting as a way of distinguishing chaos from measurement error in a data series. Nature. 1990;**344**:734-741

[17] Cheung YW, Lai KS. Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics. 1995;**13**(3):277-280

IntechOpen