

The most important question in diagnostic medicine is “Does this individual have a disease?”. The entire field of laboratory medicine and pathology has been developed to aid in answering that question. To be clinically relevant, a diagnostic test needs to be able to differentiate between the diseased and healthy state, and it also needs to be accurate and precise. Furthermore, a good diagnostic test should be clinically applicable, it should not cause harm, and in the face of ever-increasing constraints on healthcare finances, it needs to be cost-effective. In this chapter, we will address different aspects of assessing a diagnostic test [1].

The first step in assessing a diagnostic test is to examine the theoretical concept of the test and establish a causal linkage between the test and the condition of interest. The test methodology and instrument also need to be scrutinized. The precision and accuracy of the measurement instrument and test should be determined. The measurement error needs to be quantified and minimized if possible. These concepts are collectively called technical accuracy and precision.

The next step in assessing a diagnostic test is to establish the discrimination power of the test, the ability of the test to differentiate those affected by a condition from the unaffected. This step requires carefully designed clinical trials to determine the diagnostic metrics of the test and establish its accuracy. The level of diagnostic accuracy and the accuracy metrics that are used are dependent on the possible application of the test; a screening tool needs high sensitivity, while testing for a very rare condition requires high specificity. These evaluations fall under the umbrella of diagnostic accuracy [2–5].

The next critical question in assessing a diagnostic test involves determining the possible effect of the test on patient outcomes. This step involves the crucial tradeoff of benefit versus harm; the benefit of the diagnostic test should be weighed against possible adverse outcomes for the patient or the population. Also in this step, questions of applicability and feasibility should be addressed.

Finally, the cost of the new diagnostic test should also be addressed. Cost can potentially be the single most prohibitive step in adopting a new test, and to justify possible additional expenses, the cost-effectiveness of the test should be determined.

Some of the questions relating to appraisal of new diagnostic tests will be covered in Chap. 13, where we will discuss an evidence-based approach to appraisal of diagnostic studies which establish the scientific basis for new tests. In this chapter, we will start with the concept of technical accuracy and precision focusing on measurement error and statistical analysis of error. Next will be the concept of diagnostic accuracy focusing on discriminative and predictive powers of the test. Clinical impact or clinical applicability is the next step in assessing a diagnostic test. The final part of this chapter provides a brief introduction of cost-effectiveness analysis [6–8].

Technical Accuracy and Precision

Technical accuracy is the ability of a test to produce valid and usable information. Precision is essentially the reproducibility of the test, the ability to obtain very similar results if the test is repeated multiple times. Technical accuracy and precision should be determined for every new diagnostic test that is being developed, and subsequently every time a laboratory adds a test to its repertoire, it must ensure that the test is technically accurate and precise under its laboratory conditions. Evaluation of technical accuracy and precision should be an ongoing effort. Technical accuracy and precision are essentially about minimizing measurement error.

Error

Every measurement in laboratory medicine has a degree of uncertainty; this uncertainty is 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. Effectively, the results we report are best estimates of the true value.

Understanding the nature of error and quantifying it is of utmost importance in laboratory medicine as the results can have direct clinical impact on patients. High precision instruments have limited the measurement error in recent years, yet we still should estimate the error of our measurements and take corrective actions when the error surpasses an acceptable threshold.

Two main important forms of error are “random error” and “systematic error” (Fig. 2.1). The effects of systematic error and random error are additive. Random errors are caused by unpredictable changes in the measurement which may be related to instrument, sample, or environment. Addressing the causes of random error is usually difficult, and there is always a degree of random error present for every measurement.

For example, if you measure the sodium concentration of a solution with a sodium content of 140 mEq/l five separate times with results being 140 mEq/l, 141 mEq/l, 139 mEq/l, 138 mEq/l, and 142 mEq/l, then you are witnessing a

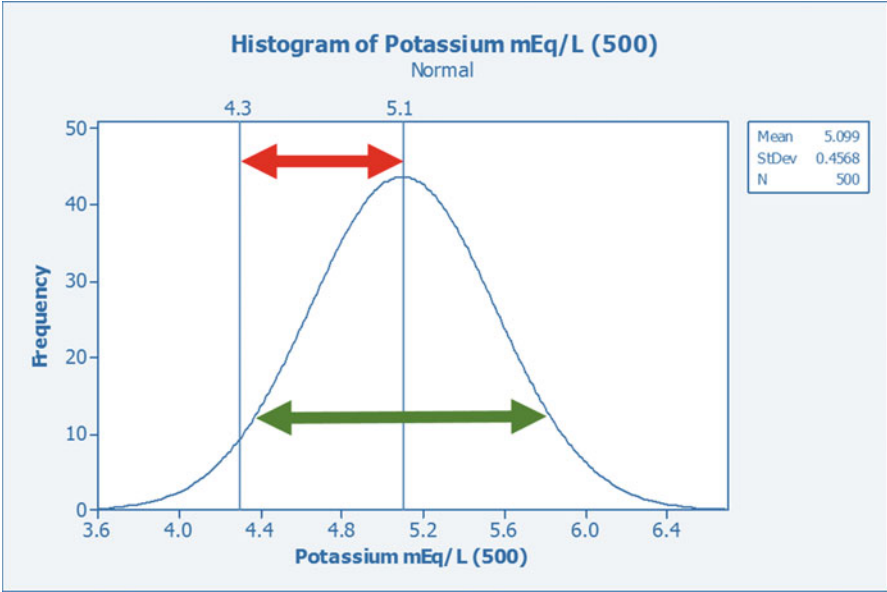


Fig. 2.1 These are the results of 500 repeated measurements of a sample with potassium concentration of 4.3 mEq/L. The *red line* shows the inaccuracy of results (systematic error), and the *green line* shows the imprecision of results (random error)

random error. The variations in results are random and cannot be predicted. However, random errors, as driven by chance, follow a Gaussian normal distribution; this allows us to use statistical analysis to quantify and address random error in our measurements. The degree of random error determines the “precision” of a test/instrument. Random error can be minimized by increasing the number of measurements. Averaging repeated measurement results is one way to report a more precise estimate of the expected value. Mean or average (\bar{x}) is the sum of measurement results divided by the number of measurements.

$$\text{Average } (\bar{x}) = \frac{x1 + x2 + x3 + x4 + x5 + \dots + xN}{N} \quad (2.1)$$

Theoretically, with infinite measurements, the mean of measurement results will be the true value. With a finite number of measurements, the true value will be within a range of the measurement mean. The range is mainly determined by the number of measurements (with more measurements the range will be narrower). This range is called the “confidence interval” and will be addressed later in this chapter.

The simplest form of random error is called “scale error.” Scale error refers to the precision of an instrument that makes a measurement as well as the precision of the reporting of the result. The measurement and reporting can be as integer numbers (i.e., 1, 2, 3, 4, ...), and a true sample value of 4.2 will be reported as

4. The 0.2-unit imprecision is due to scale error. The scale errors for instruments are determined by the resolution of measurement; higher resolution instruments will provide a more precise result. While scale error can be minimized, it can never be totally rectified as there is always a limit to the resolution of the instrument. Different tests require different resolutions and as such the scale precision differs between them. For example, in measuring cardiac troponins, a much better resolution is needed compared to measuring sodium levels. In laboratory medicine, test scales are determined by the nature of the test as well as the clinical significance of the scale. As such, the scale imprecision of tests is usually clinically insignificant. For example, a sodium level of 133 mEq/l versus 132.987 mEq/l is considered as clinical equivalents.

Systematic error is a nonzero error; averaging or repeating the results will not minimize the error. Systematic errors are reproducible and skew the results consistently in the same direction. Systematic error is otherwise known as bias. Bias can be difficult to identify and address. In laboratory medicine, the most common method of addressing bias is to use calibration. In calibration, a standard sample at different concentrations is measured, and the difference between the results and the expected value (bias) is reduced by using a correction factor. Systematic error has different causes including environmental factors, calibration problems, instrument drift, confounding factors, and lag time errors. Systematic error determines the accuracy of the results [9, 10].

Accuracy refers to the proximity of the measured value to the expected (true) value. Precision, on the other hand, deals with repeatability of the results and refers to consistency of results from repeated independent measurements. Precision is a measure of reliability and reproducibility. For each test, both accuracy and precision need to be addressed. These concepts are shown in Fig. 2.2.

There are different ways of reporting precision including fractional uncertainty and confidence interval. Fractional uncertainty is the ratio of uncertainty to the measured value. The confidence interval will be discussed later in this chapter.

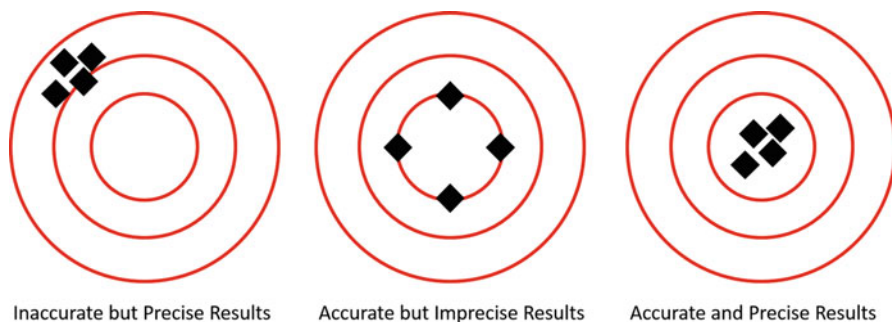


Fig. 2.2 This figure depicts the concepts of accuracy and precision

$$\text{Fractional uncertainty} = \frac{\text{Uncertainty}}{\text{Measured value}} \tag{2.2}$$

Accuracy can be reported as relative error and shows the ratio of drift to the true value. Note that the relative error is directional (can be positive or negative) with a minus relative error signifying a systematic error that underestimates the result.

$$\text{Relative error} = \frac{(\text{Measured value} - \text{True value})}{\text{True value}} \tag{2.3}$$

Standard Deviation

Standard deviation (σ) or (SD) is the uncertainty associated with each single measurement. In other words, standard deviation shows the degree of variation (spreading) of measurement results. Standard deviation is a useful measure in variables that follow a Gaussian normal distribution. Higher standard deviations signify a wider spread of data (Fig. 2.3).

Standard deviation is the square root of the variance (σ^2). Variance is the sum of squared deviation of every measurement from the mean:

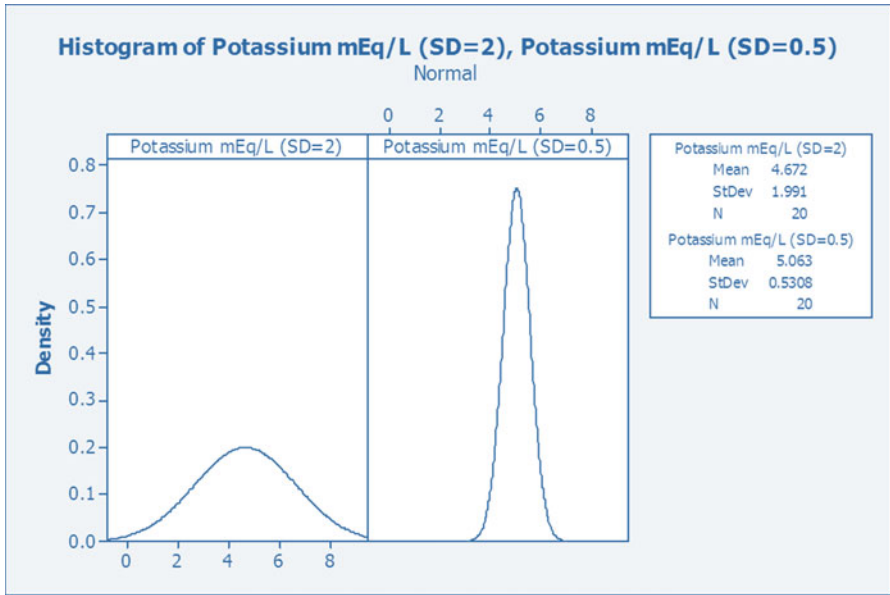


Fig. 2.3 Histogram plots of potassium measurements with a SD of 2 and 0.5. Note that as the SD increases the Gaussian bell curve becomes wider (wider spreading)

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} \quad (2.4)$$

where N is the number of measurements (or size of the sample) and (\bar{x}) is the sample mean.

Thus, standard deviation can be written as

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} \quad (2.5)$$

As standard deviation shows the spreading of uncertainty of a measurement, it is used in calculating standard deviation of mean (also known as standard error) which is in turn used to calculate the confidence interval.

Confidence Interval

Confidence interval (CI) is the range of values, estimated from sample data, which is likely to include a population parameter (Θ). In epidemiologic studies, the population parameter is usually the population mean (μ), and the sample mean (\bar{x}) is used to measure the confidence interval. In laboratory medicine, the parameter is usually the result of a test with the confidence interval being a range which is likely to include the actual measurement.

Confidence interval is centered around the measured parameter (either sample mean or test result) with the range defined by level of confidence (C). Level of confidence refers to the probability that the range contains the actual value. As the level of confidence increases, the range becomes narrower, or, conversely, the lower the level of confidence, the broader the range will be. Confidence levels are usually set at 90%, 95%, or 99%. A confidence interval of 95% means that there is a 0.95 probability that the actual value is within the range provided. For most measurements, a level of confidence of 95% is considered acceptable. Figure 2.4 shows the confidence interval on a normal density curve. For the purposes of this

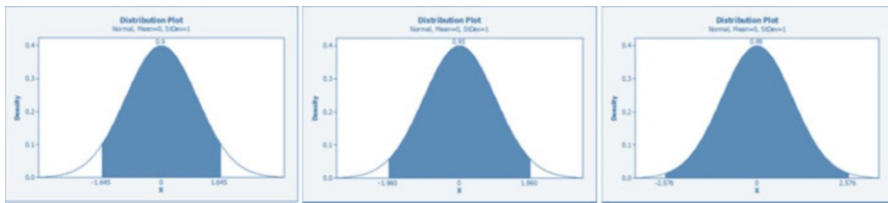


Fig. 2.4 Normal density curves depicting 90%, 95%, and 99% confidence interval of a normally distributed sample with a mean of 0 and standard deviation of 1

chapter, we assume that the measured value follows a normal distribution. Data from large sample size that does not follow a normal distribution can be approximated to a normal distribution using the central limit theorem which is discussed in the next chapter.

As the number of measurements increase (in population statistics as the sample size increases), the confidence interval will become narrower. Repeated measurements reduce the effect of random error on the mean test result thus leading to increased precision. As you will see below, confidence interval is a function of the mean (\bar{x}), confidence level, standard deviation (σ), and sample size (n). The larger the sample size, the less effect will standard deviation have on the measurement (i.e., the smaller the standard error will be). Figure 2.5 shows the effect of repeated measurements on increased accuracy of prediction.

In cases where the mean (μ) is unknown but the standard deviation (σ) is known, the formula for confidence interval (CI) is:

$$CI = \bar{x} \pm z \frac{\sigma}{\sqrt{n}} \tag{2.6}$$

The z-score is used for data that follow a normal distribution. The z-scores for 90%, 95%, and 99% level of confidence are 1.645, 1.96, and 2.576, respectively. In the potassium sample with 500 measurements, we have a sample mean of 5.099 and sample size of 500. If the standard deviation was known to be 0.5, then we can calculate the 95% CI:

$$95\%CI_{\text{potassium}(500)} = 5.099 \pm 1.96 \frac{0.5}{\sqrt{500}} = 5.055 - 5.142 \tag{2.7}$$

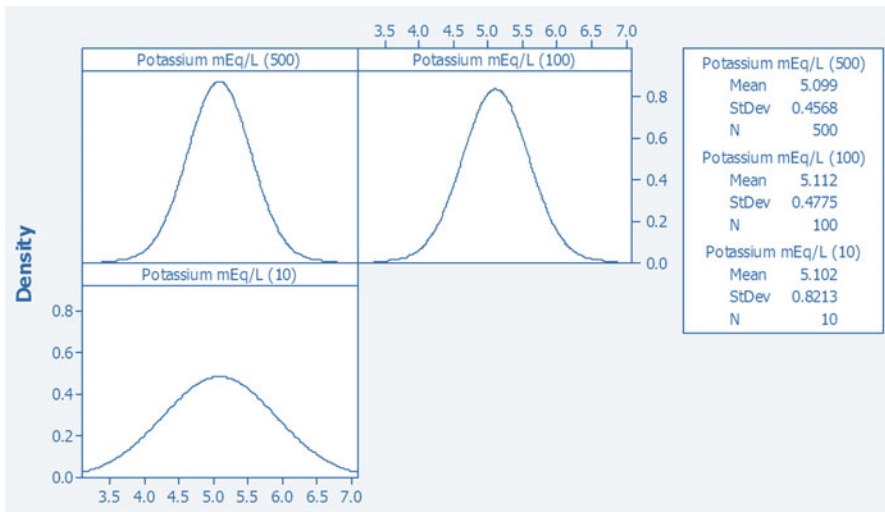


Fig. 2.5 Histograms showing repeated measurements of potassium in a 5.1 mEq/L potassium solution. As the number of measurements increases, the confidence interval becomes narrower

$\frac{\sigma}{\sqrt{n}}$ is also known as standard error of mean ($\sigma_{(\bar{x})}$). Thus, the confidence interval can be simplified as

$$CI = \bar{x} \pm z^* \sigma_{\bar{x}} \quad (2.8)$$

If the mean and standard deviation are both unknown, or if the sample size is too small (<30) for z -scores to be used, then an alternative formula is used for calculating confidence intervals. If the standard deviation is unknown, then the sample standard deviation (s) is used as an estimate of population standard deviation.

$$CI = \bar{x} \pm t \frac{s}{\sqrt{n}} \quad (2.9)$$

In these cases, the confidence level is determined by t distribution with $n-1$ degree of freedom. Thus, in the potassium sample with 10 measurements, we have a sample mean of 5.102, sample standard deviation of 0.8213, and sample size of 10. The $n-1$ t -score ($10-1 = 9$) for a 95% CI is 2.262. Subsequently, the 95%CI for the sample is:

$$95\%CI_{\text{potassium}(10)} = 5.102 \pm 2.262 \frac{0.8213}{\sqrt{10}} = 5.689 - 4.514 \quad (2.10)$$

Note that the t -scores provide a wider confidence interval compared to z -scores.

As sample size increases, the t -scores will become closer to z -scores. Also as the sample size increases, the standard deviation of the sample will be closer to the standard deviation of the population. When measuring potassium 500 times with a known standard deviation of 0.5, we observe that the sample standard deviation is 0.456, while measuring the potassium 10 times provides a sample standard deviation of 0.821. In other words, repeated measures can lead to increased accuracy [11–19].

Calculating Reference Intervals

Reference interval refers to the range of values of a measurement in healthy individuals (e.g., the range of potassium in healthy adults is 3.5–5.1 mEq/L). Reference interval is a very important information that should be provided with every quantitative test to allow the clinicians to interpret the results and determine if a patient's results are abnormal. In the next section, where we introduce diagnostic accuracy, you will see that reference interval and its overlap with values from diseased individuals has a big role in discriminative power of a diagnostic test.

If a result is within the reference range, then these results are within a certain distance of the population mean and part of normal distribution. These results are alternatively known as within normal limit (WNL). The upper and lower limits of the normal distribution of the mean are determined using the population standard

deviation. It is generally accepted that the normal limit is $\pm 2SD$ of the mean (with 95% of the healthy individual results falling within the normal limit). If a reference range is calculated in this manner, then it is called standard range. The measurements used in calculating reference ranges come from a population of healthy individuals. However, if characteristics of subgroups of the population affect the measurement, then a different reference interval should be calculated and used for each subgroup (e.g., creatinine reference range is different based on gender).

The most straightforward way to calculate a reference interval is to measure the values in a reference group of healthy individuals and sort the values from the least to the most. In this method, results that are at the 2.5–97.5% percentile (or any arbitrary cutoff) will be considered as the lower and upper limit of the reference interval, respectively. This method, despite simplicity, is not adequately reliable, and it is generally preferred that the reference interval is calculated using an arithmetic normal distribution or log-normal distribution method (discussed in Chap. 3). However, in instances where the data does not follow a Gaussian or log-normal distribution, this method can be employed.

In calculating the reference range for a variable, the assumption is that the measurements of the variable in the population follow a normal Gaussian distribution. As the population mean and standard deviation are usually unknown, then they must be estimated using a sample of the population. Using these estimates, then the 95% prediction interval (95%PI) is calculated as

$$95\%PI = \bar{x} \pm t_{0.975, n-1} \sqrt{\frac{n+1}{n}} \sigma \quad (2.11)$$

In cases where the sample size is greater than 30, the t distribution is considered as equaling 2.

For example, using a sample of 30 patients with an average potassium level of 4.5 mEq/L and standard deviation of 0.5, we can calculate the reference range for potassium as follows:

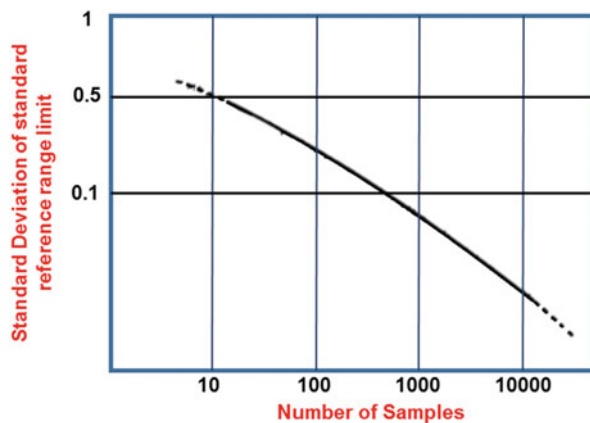
$$95\%PI = 4.5 \pm 2 \times \sqrt{\frac{31}{30}} \times 0.5 = 4.5 \pm 1.01 \quad (2.12)$$

with the upper limit of reference range being 5.51 mEq/L and the lower limit of reference range being 3.49 mEq/L.

The reference interval can have its own confidence interval. This confidence interval is dependent on the standard deviation of the standard reference interval. The size of the standard deviation is a logarithmic function of the size of the sample with larger sample leading to smaller standard deviation (Fig. 2.6).

In our previous example, the standard deviation of the standard reference interval for a sample size of 20 is 0.4 of the primary value or in other words $0.4 \times 1 = 0.4$ mEq/L. Consequently, we can estimate the 95% confidence interval of the reference range limits as

Fig. 2.6 This log-log graph shows the standard deviation of standard reference range limit versus the number of samples



$$95\%CI(\text{upper reference limit}) = 6.55 \pm 2 \times \sqrt{\frac{31}{30} \times 0.4} = 6.53 \pm 0.81 \quad (2.13)$$

$$95\%CI(\text{lower reference limit}) = 2.45 \pm 2 \times \sqrt{\frac{31}{30} \times 0.4} = 2.47 \pm 0.81 \quad (2.14)$$

These calculations are correct for all measurements that follow a Gaussian normal distribution. Goodness of fit tests such as Kolomogorov-Smirnov or Shapiro-Wilk can be employed to determine if the data has a Gaussian normal distribution.

Many laboratory tests, however, follow a log-normal distribution. One of the main reasons for this is the fact that most physiologic parameters that are measured can only assume nonnegative numbers (i.e., the results are always positively skewed). In these tests, unless the standard deviation is small compared to the mean, the Gaussian normal distribution cannot be used, and instead a log-normal distribution should be used. In other words, in measurements where the standard deviation is small compared to the mean, even if the sample measurements are positively skewed, the abovementioned calculation can still be used. As the standard deviation increases, however, log-normal distribution should be employed.

The simplest way for calculating the reference interval for a test with log-normal distribution is to calculate the natural logarithmic values of all the measurements. Consequently, arithmetic normal distribution reference interval calculations can be used to determine the lower and upper limits of the logarithmized values. The exponentiated values of these upper and lower limits will form the upper and lower limits of the reference value.

The switch to a log-normal distribution is made based on a difference ratio for the lower and upper limits. The difference ratio can be calculated as

$$\text{Difference ratio} = \frac{|\text{Limit}_{\text{Log-normal}} - \text{Limit}_{\text{Normal}}|}{\text{Limit}_{\text{Log-normal}}} \quad (2.15)$$

This difference ratio should be calculated separately for the lower limit and the upper limit. A difference ratio of more than 0.1 is considered as indicative of the need to use the log-normal distribution. The calculation of difference ratio, however, can be a cumbersome task, and thus a measure known as coefficient of variation can be used as a proxy for difference ratio. Coefficient of variation (CV) is the ratio of standard deviation to the mean.

$$\text{CV} = \frac{\sigma}{\bar{x}} \quad (2.16)$$

The lower limit of reference range is more sensitive to coefficient of variation, and a CV of 0.213 is the threshold for using a log-normal distribution for the lower limit. For the upper limit, due to positive skewedness of data, a higher CV of 0.413 is considered as the critical threshold.

In general, it is always a good idea to provide a histogram of values used for determining the reference value to the clinicians. This will allow them to better understand the reference interval [20].

Calculating Sample Size for Reference Interval Estimation

In calculating the sample size, the desired quantile of reference data (p), the desired quantile of confidence interval (α), and the desired quantile of reference interval (β) should be decided. Quantiles are defined intervals of the data; usually the data is divided into 100 quantiles each with equal number of the values. The reference interval is constructed to include the middle $\beta\%$ of the population. Usually in these calculations, α and β are set equally. After deciding these values, the corresponding z -values of Z_p , $Z_{(1-\alpha/2)}$, and $Z_{(1-\beta/2)}$ should be used to calculate the sample size. Another parameter of the formula is the relative margin of error (Δ) which is the percentage ratio of the width of the confidence interval for the reference limit to the width of the reference limit. Ideally, this margin of error should be small (i.e., the width of the CI for the reference interval limits should be small compared to the reference interval width), and usually the Δ is set at 10%. The formula for sample size calculation is as follows:

$$n \geq \frac{z_{(1-\frac{\alpha}{2})}^2 \left(D + \frac{z_p^2}{2} \right)}{z_{(1-\frac{\beta}{2})}^2 \left(\frac{\Delta}{100} \right)^2} \quad (2.17)$$

where (n) is the sample size and D is a constant which is equal to 1 if there are no subgroups in the sample (i.e., the same reference interval is used for all patients). If the test and the reference interval are dependent on a covariate, then the D value is

determined based on the nature of the covariate; with a uniformly distributed sample, D is 4. For normally distributed covariates, D is 5. For instances where the covariate can be grouped into three groups, D is $5/2$.

For example, if you want to determine the reference interval for sodium concentration, consider that you need 80th quantile of the range included in the reference range (P), and you want the alpha to be 0.05 and beta to be 0.20, and then you can calculate the sample size using the above equation (z -scores for 0.9725, 0.95, and 0.80 are approximately 1.9, 1.64, and 0.84, respectively) [21].

$$n \geq \frac{z_{(1-\frac{0.5}{2})}^2 + (1 + 0.84^2/2)}{z_{(1-\frac{0.2}{2})}^2 \times 0.1^2} \cong 156 \quad (2.18)$$

Diagnostic Accuracy and Testing for Accuracy

Diagnostic accuracy refers to a test's discrimination power that allows it to identify presence of a disease or condition in an individual. Different measures such as sensitivity and specificity are considered as proxies for diagnostic accuracy. It is important to know that diagnostic accuracy measures cover different aspects, and depending on the clinical question, a specific set of measures should be used. Furthermore, these measures are dynamic and can change per different parameters mainly population characteristics; for example, disease prevalence can affect diagnostic accuracy. Diagnostic accuracy also suffers from the "gold standard" problem, where inaccuracies in the gold standard test can confound interpretation of the diagnostic accuracy studies. Diagnostic accuracy studies usually lack statistical power or fail to follow standard procedures further complicating the issue of diagnostic accuracy. Nonetheless, clinical utility of diagnostic tests is dependent on the diagnostic accuracy of the tests. Here we will address different indicators and measures of diagnostic accuracy.

It is important to know that some accuracy measures are more concerned with discriminative power and the ability of the test to discriminate between the diseased and healthy states. Other measures are more concerned with probability estimation and provide a likelihood of diseased state based on the test result. The most important discriminative measures are sensitivity and specificity. The most commonly used probabilistic indices are positive and negative predictive values and likelihood ratio. These latter measures are highly sensitive to disease prevalence (pretest probability (see Chap. 3)). Sensitivity and specificity, however, are not affected by disease prevalence and can be carried over to different populations.

Sensitivity and Specificity

Diagnostic tests, ideally, should be able to correctly set diseased individuals apart from healthy individuals; each diagnostic test should have a discrimination power that allows for such distinction. For tests that are binary, with a distinct positive or negative outcome, the measurement of this discrimination power is straightforward with positive outcome identifying disease state (true positive) and a negative test outcome highlighting a disease-free state (true negative). For tests that return a range of values, cutoffs should be determined that will distinguish the healthy from diseased. In the most ideal setting, the results of the test will not misdiagnose an individual. However, there is always an overlap of test outcomes between healthy and diseased individuals leading to incorrect assignment of a health state to individuals. If a healthy individual is labeled as diseased by error, this is called a “false-positive outcome.” On the other hand, if a diseased individual is misdiagnosed as healthy based on the test outcome, this is called a “false-negative outcome.” These four outcomes can be displayed in a 2 × 2 contingency table (shown in Table 2.1).

Sensitivity is a measure that shows the proportion of individuals with a positive test outcome who are correctly determined to be diseased. In other words, sensitivity is the proportion of “true positives” to all diseased individuals:

$$\text{Sensitivity (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

(2.19)

Sensitivity is usually expressed as a percentage. Sensitivity is a measure of the diagnostic test’s ability to screen for a condition. Increases in the sensitivity essentially mean that the number of “false negatives” has decreased. A test with a high sensitivity will identify a significant proportion of diseased individuals. A sensitive test can thus be used to screen individuals with a condition as any “negative” test outcome is more likely to be a “true negative.” Alternatively, it can be stated that sensitive tests can be used to *rule out* the disease or condition of interest.

Specificity in contrast is a measure that determines the proportion of “true negative” individuals who are correctly determined as not having the disease, i.e., the proportion of “true negatives” to all individuals without the condition.

Table 2.1 2 × 2 contingency table showing the outcomes of the test in columns and the disease condition status in rows

	Test outcome	
	Positive	Negative
Condition positive	True positive (TP)	False negative (FN)
Condition negative	False positive (FP)	True negative (TN)

$$\text{Specificity (\%)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (2.20)$$

Specific tests are useful for *ruling in* individuals with the condition or disease of interest. As specificity increases, the proportion of healthy individuals with a “false-positive” test outcome decreases which means that a “positive” test outcome is likely to be a “true positive.”

Highly specific tests are used as confirmatory tests in a two-step diagnostic model: the first step is to employ a population-based screening test with high sensitivity followed by a highly specific test to confirm the diagnosis in individuals with a positive screening. This two-step model is preferred as tests are unlikely to be both very sensitive and very specific. Furthermore, tests with high sensitivity tend to be more affordable than highly specific tests and are thus more suited for population-level utilization. An example of the two-model is alkaloid testing where a primary test (screening) is performed using Marquis reagent spot test and the positive test results are confirmed by gas chromatography (confirmatory test). The Marquis test is fast, affordable, and easy to perform; furthermore, it has high sensitivity; all of these characteristics make it an ideal screening tool. Gas chromatography is a cumbersome and expensive test with high specificity which makes it a good confirmatory tool. Another example is the application of VDRL and RPR test for screening of syphilis infection followed by a confirmatory FTA-ABS or TP-PA test.

There usually exists a tradeoff between sensitivity and specificity. An ideal test will have a sensitivity and specificity of 100%, but such a level of accuracy is unattainable due to multiple factors including the Bayes error rate which states that there is always an irreducible error inherent in any measurement. As a test gains in sensitivity, it tends to lose specificity and vice versa. This tradeoff between sensitivity and specificity will be explained more as part of receiver operating curves (see below).

Overall accuracy is another measure that is useful and can be extracted from Table 2.1. Accuracy is a summative measure that shows the ratio of overall correct calls by the test to all the measurements made. Accuracy is one of the measures of agreement used in validating a diagnostic test and will be covered in Chap. 11.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2.21)$$

Predictive Values

Predictive values refer to the probability of having or not having the condition of interest based on the outcome of the test. Two predictive values can be extracted from the 2×2 table. First is the “positive predictive value” (PPV) which measures the probability of having the condition of interest (TP) in individuals with a positive test outcome (TP + FP).

$$\text{Positive predictive value (PPV)} = \frac{TP}{TP + FP} \times 100 \tag{2.22}$$

“Negative predictive value” (NPV) is the probability of not having the condition of interest (TN) in individuals with a negative test outcome (TN + FN).

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN + FN} \times 100 \tag{2.23}$$

Predictive values are more useful clinical measures than sensitivity and specificity as they can directly provide an estimation to the clinician of the likelihood of their patient having or not having a condition based on a positive or negative test outcome. Predictive values unlike sensitivity and specificity are affected by the disease prevalence in the population and as such cannot be transferred from a population to population. The effect of disease prevalence on PPV and NPV is different; as the prevalence increases, the PPV increases (because the probability of having a false-positive result decreases), while NPV decreases. If the prevalence decreases, the reverse will be true; NPV will increase and PPV will decrease. The effect of prevalence is more significant on PPV than on NPV. For diseases with a low prevalence, a test with high specificity (low false-positive rate) is needed to have an acceptable positive predictive value (Fig. 2.7).

Other accuracy measures can also be extracted. A summary of these measures is shown in the table below (Table 2.2).

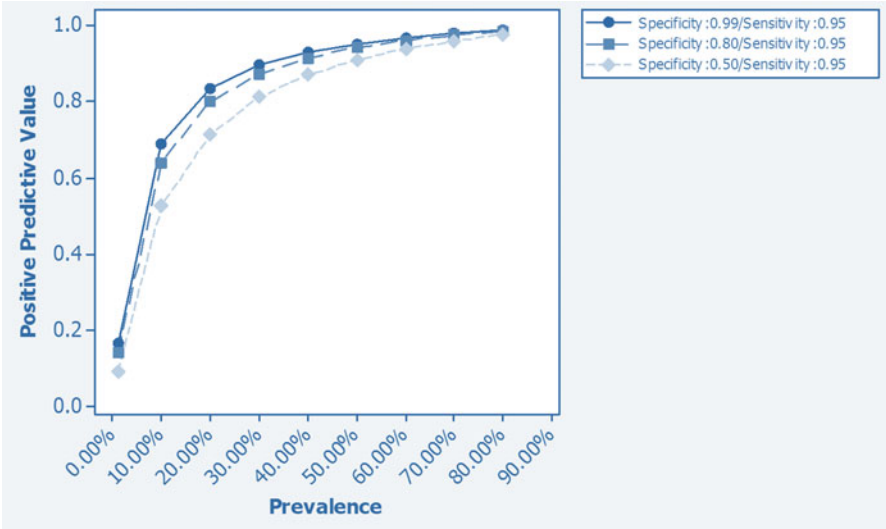


Fig. 2.7 Scatter plot of prevalence versus positive predictive value (PPV) for different test specificities. Note the marked decrease in PPV as prevalence falls below 10%

Table 2.2 Summary of diagnostic accuracy measures

	Test outcome		
	Positive	Positive	
Condition positive	<i>TP</i>	<i>FN</i>	Sensitivity = $TP / (TP + FN)$ False-negative rate (FNR) = FN / TP
Condition negative	<i>FP</i>	<i>TN</i>	False-positive rate (FPR) = FP / TN Specificity = $TN / (TN + FP)$
Accuracy = $(TP + TN) / (TP + TN + FP + FN)$	Positive predictive value (PPV) = $TP / (TP + FP)$	False omission rate (FOR) = $FN / (TN + FN)$	Diagnostic odds ratio (DOR) = $(TP \times TN) / (FP \times FN)$ = LR+ / LR- Positive likelihood ratio (LR+) = sensitivity / FPR Negative likelihood ratio (LR-) = FNR / specificity
	False discovery rate (FDR) = $FP / (TP + FP)$	Negative predictive value (NPV) = $TN / (TN + FN)$	

Receiver Operating Characteristic Curve

The basic assumption for every diagnostic test is that the diseased individuals will have different test outcomes compared to the unaffected population. Many tests return a quantitative range of values instead of a binary “positive” or “negative” value. In these tests, cutoff values must be determined that will set apart the affected from unaffected. Determining cutoff values will depend on the distribution of the values among unaffected and diseased individuals as well as the desired sensitivity and specificity levels. In a perfect test, the outcome values for the affected and unaffected population will have no overlap. There is, however, always a degree of overlap between the two populations making the decision of a cutoff value very important as different cutoff values will lead to different sensitivity and specificity levels (Fig. 2.8).

Receiver operating characteristic curve (ROC) is the graphical illustration of true positive rate (sensitivity) as the Y-axis and false-positive rate (1-specificity) as the X-axis. ROC curve is generated by plotting the cumulative distribution of sensitivity as a function of cumulative distribution of false-positive rate. Consequently, the ROC curve shows the trade-off between sensitivity and specificity. The basic concept behind the ROC curve is that a reference or test variable (test outcome) is used to classify subjects and classification performance is compared with a classifier variable (gold standard), and at different cutoff values for the test variable, true positive rate and false-positive rate are calculated and plotted as Y-axis and X-axis, respectively.

The ROC curve is usually plotted using a nonparametric generalized linear model. The most common approach was proposed by Tosteson and Begg. In the simplest form of their model, the only classifier is the indicator of the true disease status (χ_1). The other assumption will be that for test outcome values of r_1 through r_j , the subject is classified as negative (T-), and for values greater than r_{j+1} , the subject is classified as positive (T+). Subsequently, for the cumulative probabilities

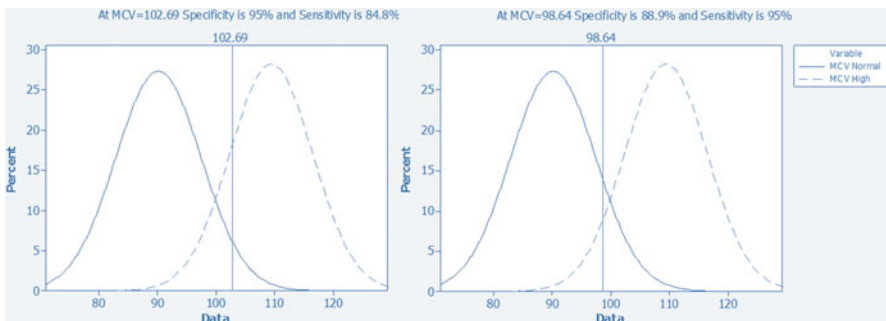


Fig. 2.8 Distribution of mean corpuscular volume (MCV) of normal population and macrocytic anemia (with an assumed 50% prevalence). If the cutoff is set at 102.69 fL, the specificity will be 95% and sensitivity will be 84.8%. Lowering the cutoff will increase the sensitivity and reduce the specificity (at 98.64 fL the sensitivity and specificity will be 95% and 88.9%, respectively)

of response, $\gamma_j(\chi_1)$, determine the response categories (TP, TN, FP, and FN). In this setting, $\gamma_j(0)$ is the probability that an unaffected individual has in a test value outcome of between r_1 and r_j (i.e., test outcome value lower than the cutoff value). This probability represents the “true negative rate” or “specificity” and consequently $1 - \gamma_j(0)$ represents the “false-positive rate” which forms the X -axis of the ROC space. $\gamma_j(1)$ will be the probability that an affected individual has in a test outcome value of between r_1 and r_j (false-negative rate), and thus $1 - \gamma_j(1)$ will be the “true positive rate” or “sensitivity” which forms the Y -axis of the ROC space. The ROC curve will be constructed by plotting all the pairs of $1 - \gamma_j(0)$ and $1 - \gamma_j(1)$ for each of the test outcome cutoff points (θ_j). The following generalized linear model will form the ROC curve:

$$g[\gamma_j(\chi)] = \frac{\theta_j - \alpha'\chi}{\exp(\beta'\chi)} \quad (2.24)$$

$$j = 1, \dots, j-1$$

α' and β' are regression parameters of location and scale, respectively. These two parameters will determine the shape of the curve and in the simplest form are defined as constants that will provide the curve a concave appearance. To make the curve smooth, smoother functions known as link functions are introduced to the generalized linear model. The most common link function used is the probit link which is based on the standard normal cumulative function, Φ . The generalized linear model with the link function applied will be:

$$\Phi^{-1}[\gamma_j(\chi)] = \frac{\theta_j - \alpha'\chi}{\exp(\beta'\chi)} \quad (2.25)$$

$$j = 1, \dots, j-1$$

ROC curve has been shown to have a nonparametric interpretation like the Mann-Whitney U test (see Chap. 6). This means that, while the distribution of test values in the affected and unaffected population usually follows a binormal distribution, other non-normal distributions can also be used in constructing a ROC curve.

ROC space is a square with X - and Y -axis range of 0–1. A diagonal line connects the top right corner of the space to the bottom left corner. This line is called the line of no discrimination and depicts a complete random association of the test variable with the classifier. The perfect classification point in the ROC space (100% specificity and sensitivity) lies at the top left corner of the space. The further the ROC curve moves away from the diagonal line toward the top left corner, the better the classification properties of the test variables will be (Fig. 2.9).

For determination of the cutoff value, two approaches can be undertaken. In the first approach, a decision must be made on the optimal level of sensitivity and specificity on the ROC curve, and the cutoff value extracted from the table of curve coordinates which shows the corresponding test value for each curve coordinate. This manual search for the cutoff value allows for choices such as choosing a cutoff for screening (high sensitivity) or for confirmation (high specificity) (Fig. 2.10).

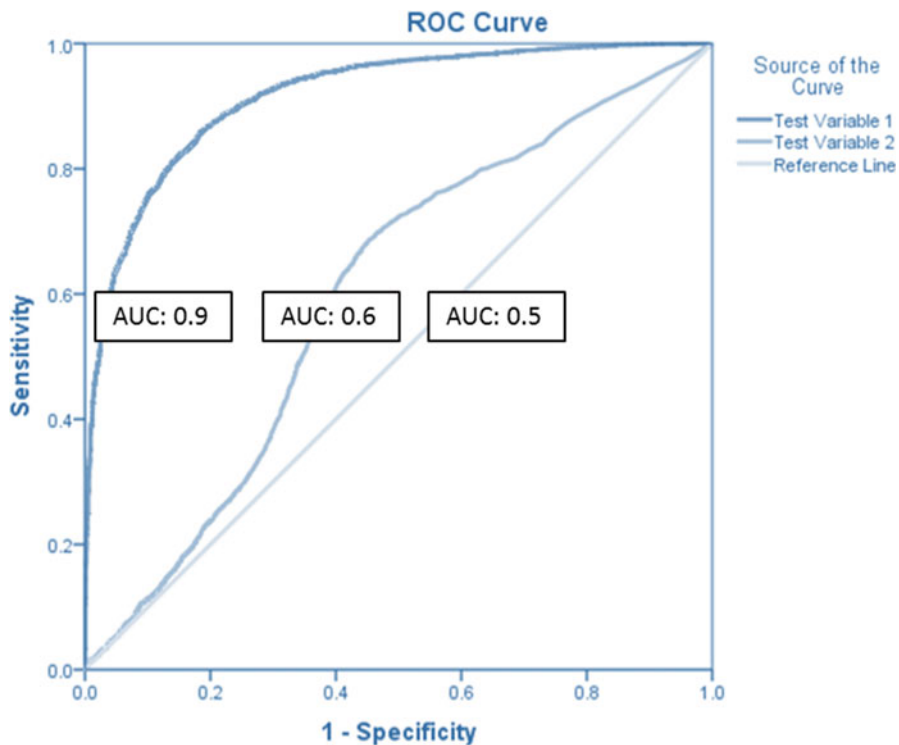


Fig. 2.9 ROC curves for two test variables are depicted. As the curve nears the *top left* corner of the ROC space, the classifying power of the test variable increases. In this figure, note that the test variable 1 is a far better classifier compared to test variable 2

If sensitivity and specificity are given equal weights, then a ROC curve analysis can be employed to determine the optimal cutoff value. Several methods of ROC curve analysis have been established. One of the oldest and simplest methods is called the Youden's index. This index is calculated as the difference of the sum of sensitivity and specificity from 1.

$$\text{Youden's index} = (\text{Sensitivity} + \text{Specificity}) - 1 \quad (2.26)$$

Youden's index can assume values between 0 and 1 with 0 showing poor diagnostic accuracy and 1 showing a perfect diagnostic accuracy (sensitivity and specificity of 100%). In ROC curve analysis using the Youden's index, the index is calculated for every coordinate on the ROC curve, and the corresponding test value for the point of maximum Youden's index is then set as cutoff value. The cutoff value set in this approach balances the sensitivity and specificity. Essentially Youden's index determines the cutoff to be at the point where the two distribution of test outcome values of the affected and unaffected population meet. Financial considerations and cost can also be criteria for determining cutoff values and can be incorporated in ROC curve analysis.

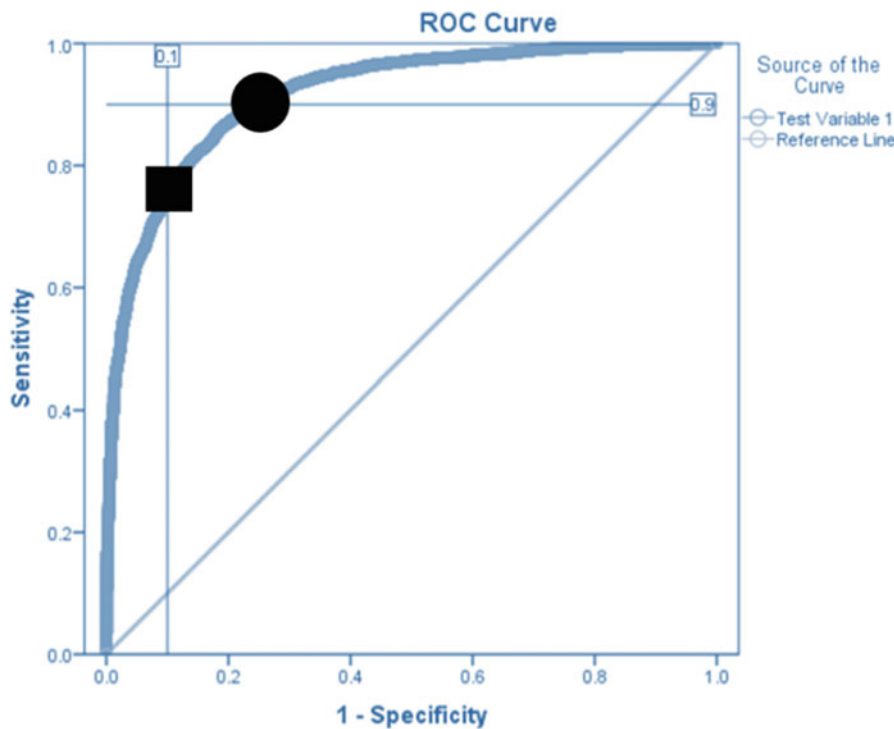


Fig. 2.10 Different cutoff values can be chosen based on the desired level of specificity and sensitivity. The square depicts a point where specificity is 90%. The circle depicts a point where sensitivity is 90%

One of the benefits of ROC curves is the ability to calculate “area under the curve” (AUC). AUC is one of the most useful measures of diagnostic accuracy and discrimination power of a test. A perfect AUC will have a value of 1; this will signify that there exists a cutoff point in test outcome values where the sensitivity and specificity will be 100%. Consequently, as AUC nears 1, the classification (discrimination) power of the test will increase. AUC can be stated in form of the probability that a randomly selected affected individual will have a higher test outcome value than randomly selected unaffected individual.

A rough estimate of levels of discrimination power based on AUC is provided in the Table 2.3 [22, 23].

Calculating AUC

The simplest approach for calculating AUC is using the trapezoidal rule. In this approach, the space under the curve is transformed to a series of rectangles and triangles, and their cumulative area is calculated. If a single cutoff point (j) is used,

Table 2.3 Levels of discrimination power based on AUC

Area under curve (AUC)	Discrimination power
0.9–1	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Acceptable
0.5–0.6	Bad
<0.5	Not useful

then the AUC calculated using the trapezoidal method will equal to $\frac{1}{2}$ ($\text{Sensitivity}_j + \text{Specificity}_j$). This method will always underestimate the true AUC.

Several nonparametric approaches can be used for better estimation of AUC. One of the methods suggested by Hanley and McNeil is called the “Wilcoxon area estimate.” In this method, due to inherent similarity between U statistics of a Mann-Whitney test and AUC, the area under curve is calculated using a rank-sum Mann-Whitney U test.

$$\text{AUC} = \frac{n_o n_1 - U}{n_o n_1} \quad (2.27)$$

where n_o and n_1 represent the sample size of the unaffected and the affected populations. In this calculation, the U statistics is calculated using the rank sum of the unaffected population (R):

$$U = R - \frac{1}{2} n_o (n_o + 1) \quad (2.28)$$

The standard error of the AUC estimation by the Hanley method can also be calculated.

$$\text{SE}(\text{AUC}) = \sqrt{\frac{\text{AUC}(1 - \text{AUC}) + (n_1 - 1)(Q1 - \text{AUC}^2) + (n_o - 1)(Q2 - \text{AUC}^2)}{n_o n_1}} \quad (2.29)$$

$$Q1 = \frac{\text{AUC}}{(2 - \text{AUC})} \quad (2.30)$$

$$Q2 = \frac{2\text{AUC}^2}{(1 + \text{AUC})} \quad (2.31)$$

While AUC is a relatively simple accuracy measure, recently there have been arguments against using it as a measure of classification power. This is because AUC is a summary measure that includes both relevant and irrelevant parts of a curve; the performance of the test at the extremes of the curve (where specificity will be very high, but sensitivity will be very low or vice versa) is usually not of

interest to the clinicians. Furthermore, AUC gives equal weight to sensitivity and specificity and may not be useful in instances where one of the measures is of greater importance.

Clinical Applicability

Establishing technical accuracy is the first step in appraisal of new tests. The next step will be assessment of diagnostic accuracy. Yet, perhaps even before this step, it is necessary to consider the clinical context in which the test will be used. Most new tests will have a similar test or diagnostic method in the clinical pathways or decision-making algorithms. There are exceptions to this rule, mainly when new screening tests are developed. Thus, an effort must be made to determine the pathway or decision-making algorithm to which the new test belongs. After the pathway is determined, we should identify the possible role of the test in the clinical pathway; a test can be used to screen for or diagnose a disease, it can also be used to guide treatment choices, or it can provide prognostic information. Some tests will also have community or population level indications; for example, they can show the genetic predisposition of the patient's offspring or can determine infectious disease carrier status of a patient.

A new test will either be upstream of a clinical pathway or have a role in "triaging" patients; it may replace an existing test in the clinical pathway or be an add-on to the existing diagnostic pathway. The decision of where the new test will be in regard to the clinical pathway will determine the characteristics and quality metrics of the test. Tests can also have non-diagnostic applications such as monitoring and prognostication.

If a test is to be a substitute for an existing test, it needs to improve upon one or few of the existing test's qualities such as accuracy, cost, harm, ease of performance, etc. Thus, to establish the superiority of the new test (or non-inferiority when factors such as cost or harm are improved in the new test), comparative studies should be conducted to gauge the new test against the existing test. Triage tests or screening tests need to be noninvasive, easy to perform, and cheap; they also need to have high sensitivity. Again, these tests need to be evaluated using clinical trials in order to establish their diagnostic accuracy. Add-on tests will help to further categorize patients in clinical pathways or determine prognosis or treatment options. These tests require higher specificity and are usually time and resource intensive to perform. Currently, a consistent proportion of new tests are focusing on add-on tests as the market for add-on tests is more targeted with less direct competition.

Ideally, the patient outcomes and quality metrics of new diagnostic tests should be measured using well-designed blinded randomized clinical trials. Other trial designs such as controlled trials, before-after studies, and prospective cohorts are also of limited use in estimating the impact of new diagnostic tests. Another option is to determine the effects of the new test on patient outcome via assessment of changes in physicians' intentions for treatment and management. However, given

the current fast pace of innovation, in certain circumstances “modeling” can be used to estimate the impact of the test. We will explore the diagnostic studies in Chap. 12 and data modeling in Chap. 15.

In assessing the clinical benefit of the tests, it is important to identify objective patient outcome measures that are affected by the diagnostic test. In many situations, finding these outcomes is problematic as a direct causative link between diagnosis and patient outcome may be lacking. The effects of a test may not just be physical but also emotional, behavioral, cognitive, or social. Furthermore, a multitude of confounding factors can obscure the true impact of the test. Lack of a targetable outcome with possible treatment options is serious argument against a new test; for example, tests that identify Alzheimer’s disease in very early stages are of limited or no clinical use as currently there are no viable treatment options available to the patients.

Secondly, possible trade-offs of utilizing the new test should be identified and balanced. The new test may be associated with direct harm due to invasiveness of the test or administration of possibly toxic or hazardous elements to the patient. Sometimes, the harm can be secondary, for example, a screening test that has good sensitivity yet poor specificity may lead to high false-positive rates which can be subjected to harmful tests or treatments as a follow-up to the screening test. As with benefits, the harm of the test should be identified and measured.

Clinical utility of the test will be determined by comparing the benefit with the harm. While there are objective methods of weighing the benefit versus harm, sometimes a subjective judgment by consensus panels of experts is needed to decide the utility of new tests.

One of the ways to assess the clinical utility of a test is to determine the absolute difference (ΔP) between pretest and posttest probabilities (see Chap. 3). This difference is dependent on test characteristics such as likelihood ratio as well as the pretest probability: as the pretest probability decreases (e.g., prevalence decreases), the likelihood ratio of the test should increase.

$$\text{Absolute difference } (\Delta P) = |\text{pretest probability} - \text{post-test probability}| \quad (2.32)$$

Absolute probability difference is sometimes difficult to interpret; in reality, the utility of a clinical test is to allow clinicians to alter the care and management of a patient thus providing benefit and avoiding harm. Further calculations are needed to extract the net benefit of a test:

$$\text{Net benefit} = (\Delta P \times r_i \times (b_i - h_i)) - h_t \quad (2.33)$$

where r_i is the rate of changes in the interventions based on probability changes (e.g., to follow curative treatment versus palliative treatment), b_i is the benefit of the changes in the interventions to the patient, h_i is the harm of the changes in the interventions to the patient, and h_t is the harm associated with the test itself. Cost can also be considered in this formula.

The question of clinical utility is usually addressed and assessed by governance and supervisory entities (such as FDA or CLIA) or the manufactures. For the practicing pathologist, it is often more important to be able to evaluate these studies and advisories (see Chap. 13) and decide on the issue of clinical applicability or relevance in their own practice setting. Clinical relevance is determined by answering two questions: transferability and feasibility.

Transferability

Studies for determining the diagnostic and technical accuracy of tests are usually performed in controlled setting with limitation on patient population, biases, and confounding factors. The study setting, as result, can potentially considerably different from clinical setting. Clinical settings can also vary across geographical or community spectra. Nonetheless, the pathologist needs to determine whether test metrics can be transferable to his/her own practice setting.

One of the determinants of transferability is the patient spectrum. It is important that the spectrum of patients seen in the pathologist's practice match or be close to the spectrum of trial subjects. Sometimes, trials upon which diagnostic accuracy is established by manufactures suffer from "spectrum bias" where highly selected patients with controlled parameters are included in the study. Outcome of such studies may not be generalizable, and adopting those tests to a different practice setting with a different population metrics may be problematic. In these setting, before a test is adopted, test validation is required, where the performance of the test is assessed in a representative sample of the population and the results are compared with the test developers' study results. Existence of gold standard tests can help in validating a new diagnostic test (see Chap. 11).

Sensitivity and specificity are thought to be independent of disease prevalence, yet it has been established that in highly controlled study samples with stringent inclusion and exclusion criteria, the calculated specificity and sensitivity may differ from clinical practice, especially in early adaptation settings of a new test due to "indication creep," whereby clinicians order the test for increasingly broad indications, the case-mix, and definition of affected individual changes.

Transferability is also an issue when the test is to be used in a different role than originally anticipated. For example, EGFR status is tested in stage IV lung adenocarcinomas, and the test validation has been performed for that setting. If EGFR status is tested in all lung cancer patients irrespective of stage or type, then a shift in the role has occurred. While, in theory, there can be justification for this transfer, the evidence to support it is lacking.

Another issue in transferability is regarding cross platform and technology transfer. As assays, platforms, technologies, and even test version are changed, there may be a need for revalidation of the test unless there is enough evidence to support the cross validation of these factors.

Feasibility

Determining feasibility of performing a test depends on the practice setting. The needs of the population served and the resources at the disposal of the lab as well as the requirements of the test determine the feasibility of adopting a new test or platform. In smaller setting, answers to these questions can be easier to find, yet in large practice settings, often, a feasibility study is needed.

A feasibility study needs to answer the following questions:

- Will the pathologists, clinicians, patients, and technicians accept the new test? (Acceptability)
- Will there be enough demand for the new test to justify the capital investment, retraining of staff, and the recurring costs? (Demand)
- Will performing the test be possible in the current setting with the available resources? (Implementation)
- Will the test be practical in the practice setting? Is the level of complexity or cost acceptable in the practice setting? (Practicality)
- Can the test be adapted and changed to fit the lab and the population served by the lab? (Adaptability)
- Can the test be integrated into the current lab routines and systems? (Integration)

Feasibility studies need stake holder analysis with participation of the lab director, clinicians, and technicians. Need assessment may sometimes be needed if the demand for a new test is not clear-cut; this assessment is usually in form of practice surveys of the clinicians. Cost analysis and breakdown of cost burden of the new test will also be necessary in evaluating the feasibility of adopting the test. Finally, small-scale runs or pilots can be helpful in better understanding the feasibility of adopting the new test or platform.

It must be noted that clinical utility is a continuous and ongoing issue which needs to be periodically revisited. As the clinical practice and technologies evolve, it may be necessary for the lab to adapt and change, and this requires constant review of the current tests and possible expansions or innovations that can improve, surpass, or replace the current tests [24].

Cost-Effectiveness Analysis

When adopting a new diagnostic test, the final yet perhaps one of the most critical questions will be the cost. The lab directors are eventually responsible for the financial state of the lab and need to decide if adopting the new test will be financially viable. In a purely financially driven setting, outcomes are often ignored, and the entire enterprise is set up to minimize cost and maximize profit. However, in most laboratories, improved patient outcome is the ultimate goal, and thus cost should be considered alongside effectiveness or benefit; if an improved outcome is attainable even at a higher cost, this may justify the cost.

Cost can be broken down to capital and recurrent costs; capital is the financial resources dedicated to procure the equipment, and the recurring costs are the financial resources required for continued operation of the equipment and running of the test (including reagents, controls, and labor). Capital cost should always be weighed against discounting; due to inflation, the true value of an investment made today will be discounted in the future; thus, in calculating investment returns, capital adjustment is needed. The costs should be summarized in form of a per-test cost: the financial resources needed to run a single test. In calculations of cost-effectiveness, we will use the per-test cost as the measure of cost.

Often a test will replace an existing test. In these settings, the per-test costs of the two tests will be directly compared, or, alternatively, a measure called “incremental cost” can be used which is the difference in the per-test cost of the new test versus the old test.

Measuring benefit can be more difficult than measuring cost; crude measures such as life expectancy or changes in mortality can be used, but these measures don’t encompass all relevant aspects of patient outcome. Subjective measures such as visual analogue scales can also be used, but they often vary considerably from patient to patient and can be difficult to interpret. Summary measures such as “disability-adjusted life years” (DALY) or “quality-adjusted life years” (QALY) are better suited for measuring effectiveness and are more relevant to patient outcomes.

In measuring both cost and benefit (outcome), the “perspective” should also be considered: whether the cost and outcome are measured at lab level, institution level, or social level. This is a fundamental decision that can make tests that appear cost-ineffective at lab level highly cost-effective at social level (e.g., using advanced nuclear amplification detection methods for screening of tuberculosis instead of smears). Further discussion of the measuring units of costs and effectiveness is beyond the scope of this book.

In comparing a new test (N) versus an existing test (O) and knowing the costs and effectiveness of each test, then cost-effectiveness can simply be stated as

$$\begin{aligned} &\text{Incremental cost – effectiveness ratio} \\ &= \frac{\text{Cost}_N - \text{Cost}_O (\Delta \text{Cost})}{\text{Effectiveness}_N - \text{Effectiveness}_O (\Delta \text{effectiveness})} \end{aligned} \quad (2.34)$$

A perfect test should have a negative “ Δ cost” and a positive “ Δ effectiveness.” A test in which the effectiveness decreases and the cost increases will also be automatically rejected. In cases where the changes in cost and effectiveness are contradictory, the decision will be based on core values of the lab: cost minimization versus patient outcome maximization. One way of measuring effectiveness is to use “posterior odds” (see Chap. 3). Posterior odds are products of prior odds (calculated using disease prevalence) and likelihood ratio.

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio} \quad (2.35)$$

Alternatively, in the formula for incremental cost-effectiveness ratio (ICER), the net benefit of the test (see above) can be used in place of Δ *effectiveness*.

“Decision analytical model” will often provide more insight into cost-effectiveness analysis (Fig. 2.11). This model follows two steps, with the first step involving calculating the “hypothetical performance and cost” of the new test, and if the new test passes this hypothetical step, then a clinical trial is undertaken to calculate the actual cost considerations of the new test. If actual test diagnostic accuracy data and cost estimations are available, then the first step can be skipped, and the cost-effectiveness is determined using the second step.

For example, assume that we are constructing the decision analytical model for a disease with a prevalence of 0.01. The sensitivity and specificity of the current test (T_0) are 80% and 85%, respectively, and the sensitivity and specificity of the new test (T_1) are 90% and 90%, respectively. The cost of the current test is 5\$ and the cost of the new test is 20\$.

In this example, we are employing a health system perspective, and we calculate the costs as the total cost burden of the health system. Based on this, a true positive result will cost 5000\$ (early treatment) plus test cost and lead to a DALY of 0.1. A false-negative result will lead to a cost of 10,000\$ (late treatment) plus test cost and lead to a DALY of 10. A true negative result will lead to only the test cost and a DALY of 0. The false-positive result will lead to a cost of 5000\$ (early treatment) plus test cost and lead to a DALY of 0.1.

Now we can construct the model (Fig. 2.12). As you can see, despite the higher cost of the new test, at health system level, employing this test will lead to both cost saving and reduction of average DALYs. The results of this cost-effectiveness analysis support the adoption of the new test in place of the current test [25–29].

Summary

We have shown that to assess a diagnostic test, a stepwise approach is needed. The first step of the assessment is technical assessment. In this step, scientific and technical issues related to the test are evaluated, and possible sources of error are identified and addressed. It is important that pathologists know technical test parameters especially precision and accuracy. The pathologist should be able to evaluate the scientific merit of the body of evidence supporting a new test. We will discuss this at length in Chap. 12, where we provide an approach for critical appraisal of literature.

The next important question is to determine the diagnostic accuracy of the test, in other words, to determine if the test can measure what it was designed to measure and whether it has enough discrimination power to be clinically relevant.

This is followed by a closely related step, in which the clinical applicability of the test is assessed: the clinical benefit of the test should be determined and the pathologists needs to assess whether the test is transferable to his/her setting and if so, whether it is feasible to implement the new test or not.

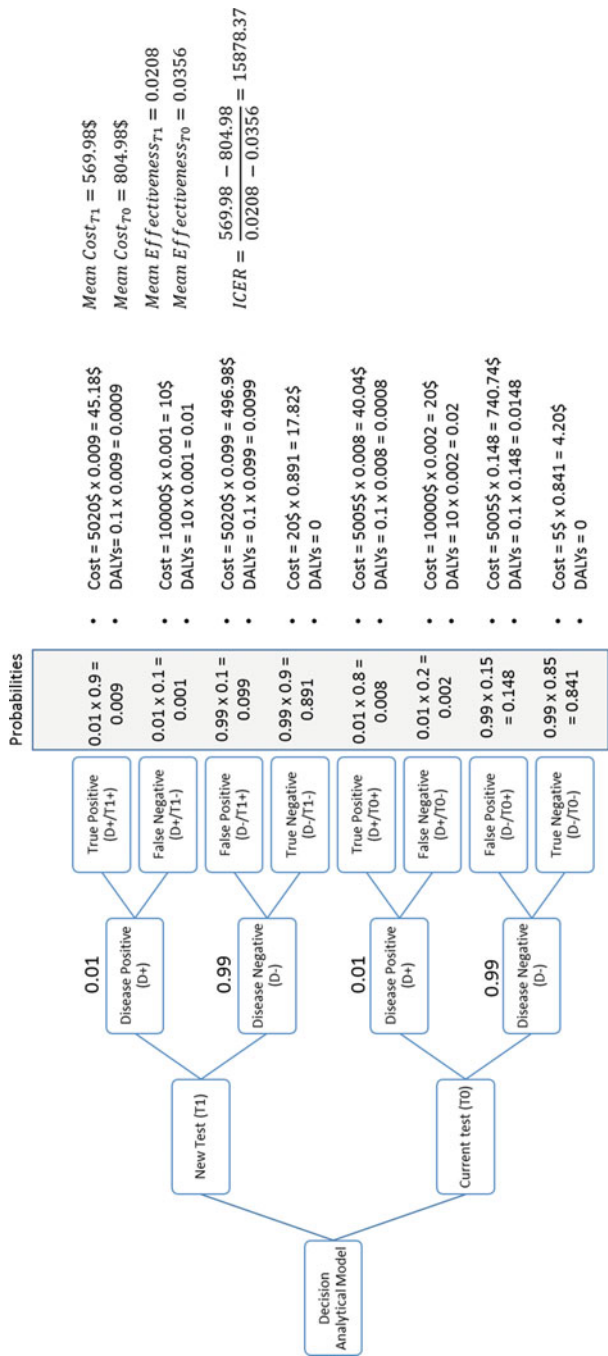


Fig. 2.12 Decision analytical model for the example provided

The final yet critical assessment is to assess the cost of adopting the new test and whether the incremental cost over existing tests is justifiable. Cost-effectiveness analysis is a power tool with which every lab director should be familiar.

We will revisit some of the concept introduced in this chapter further in the book, specifically, in Chap. 10 where we talk about test validation.

References

1. McPherson RA, Pincus MR. Henry's clinical diagnosis and management by laboratory methods. Elsevier Health Sciences: USA; 2016.
2. Crook M. Clinical governance and pathology. *J Clin Pathol.* 2002;55(3):177–9.
3. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. *J Clin Epidemiol.* 2007;60(11):1116–22.
4. Garfield S, Polisen J, Spinner DS, Postulka A, Lu CY, Tiwana SK, Faulkner E, Poullos N, Zah V, Longacre M. Health technology assessment for molecular diagnostics: practices, challenges, and recommendations from the medical devices and diagnostics Special Interest Group. *Value Health.* 2016;19(5):577–87.
5. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. *BMC Med Res Methodol.* 2013;13(1):12.
6. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. *Br Med J.* 2002;324(7335):477.
7. Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, Smith PG, Sriram N, Wongsrichanalai C, Linke R, O'Brien R. Evaluation of diagnostic tests for infectious diseases: general principles. *Nat Rev Microbiol.* 2008;8:S16–28.
8. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* 1988;240(4857):1285.
9. Bonini P, Plebani M, Ceriotti F, Rubboli F. Errors in laboratory medicine. *Clin Chem.* 2002;48(5):691–8.
10. Acken JM, Millman SD. Fault model evolution for diagnosis: accuracy vs. precision. In: *Proceedings of the custom integrated circuits conference*; 1992. p. 13–4.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, De Vet HC, Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem.* 2003;49(1):7–18.
12. Jennings L, Van Deerlin VM, Gulley ML. Recommended principles and practices for validating clinical molecular pathology tests. *Arch Pathol Lab Med.* 2009;133(5):743–55.
13. Šimundić AM. Measures of diagnostic accuracy: basic definitions. *EJIFCC.* 2009;19(4):203.
14. Yuoh C, Elghetany MT, Petersen JR, Mohammad A, Okorodudu AO. Accuracy and precision of point-of-care testing for glucose and prothrombin time at the critical care units. *Clin Chim Acta.* 2001;307(1):119–23.
15. Sirota RL. Error and error reduction in pathology. *Arch Pathol Lab Med.* 2005;129(10):1228–33.
16. Zarbo RJ, Meier FA, Raab SS. Error detection in anatomic pathology. *Arch Pathol Lab Med.* 2005;129(10):1237–45.
17. Plebani M. The detection and prevention of errors in laboratory medicine. *Ann Clin Biochem.* 2010;47(2):101–10.
18. Bossuyt PM, Irwig L, Craig J, Glasziou P. Diagnosis: comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006;6:1089–92.
19. Apple FS, Jesse RL, Newby LK, Wu AH, Christenson RH. National Academy of Clinical Biochemistry and IFCC Committee for Standardization of Markers of Cardiac Damage

- Laboratory Medicine Practice Guidelines: analytical issues for biochemical markers of acute coronary syndromes. *Circulation*. 2007;115(13):e352–5.
20. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results. *Am J Clin Pathol*. 2010;133(2):180–6.
 21. Bellera CA, Hanley JA. A method is presented to plan the required sample size when estimating regression-based reference limits. *J Clin Epidemiol*. 2007;60(6):610–5.
 22. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. *Med Decis Making*. 1988;8(3):204–15.
 23. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627.
 24. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, Bakken S, Kaplan CP, Squiers L, Fabrizio C, Fernandez M. How we design feasibility studies. *Am J Prev Med*. 2009;36(5):452–7.
 25. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. *Med Decis Making*. 2009;29(5):E22–9.
 26. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. *JAMA*. 1996;276(16):1339–41.
 27. Mushlin AI, Ruchlin HS, Callahan MA. Costeffectiveness of diagnostic tests. *Lancet*. 2001;358(9290):1353–5.
 28. Brunstein J. Cost-effectiveness considerations with molecular diagnostics in oncology. *MLO Med Lab Obs*. 2016;48(5):30.
 29. Gray AM, Clarke PM, Wolstenholme JL, Wordsworth S. *Applied methods of cost-effectiveness analysis in healthcare*. Oxford: OUP; 2010.

Introduction to Statistical Methods in Pathology

Momeni, A.; Pincus, M.; Libien, J.

2018, XIV, 317 p. 84 illus. in color., Hardcover

ISBN: 978-3-319-60542-5