Received 12.26.05 | Revisions Received 1.5.06 | Accepted 1.6.06

Statistical Methods for Establishing and Validating Reference Intervals

Roger L. Bertholf, PhD

(University of Florida Health Science Center/Jacksonville, Jacksonville, FL)

DOI: 10.1309/CBMHPRFNLU1XA4XV

Abstract

Reference intervals are an essential part of laboratory medicine, and accreditation standards require that every laboratory result is accompanied by an appropriate reference interval to provide guidance in the interpretation of the test. Furthermore, the Clinical Laboratory Improvement Act of 1988 (CLIA '88) requires laboratories to verify that the reference interval accompanying a laboratory result is appropriate for the patient population the laboratory serves. These 2 tasks are fundamental to providing quality laboratory services. In this review, we will consider the statistical methods that can be applied to establish and validate reference intervals.

After reading this article, the reader should understand the statistical components behind establishing reference intervals.

Generalist exam 90601 questions and corresponding answer form are located after the CE Update section on p. 311.

Reference intervals delimit the expected results of various laboratory tests in healthy individuals, and provide some guidance in the interpretation of patient results. It can be difficult, however, to define the appropriate reference interval for any particular laboratory test. Many factors, including age, gender, race, posture during specimen collection, geographical location, diurnal variations, and even seasonal changes may influence the results of laboratory tests.[1] These factors are partially responsible for the intra- and inter-individual variations observed in the results of laboratory tests, and reference intervals should reflect these variations. Reference populations that are selectively enriched with individuals predisposed to higher or lower values will result in a significantly biased reference range. A reference population should be composed of healthy individuals who are demographically matched to the patient population the laboratory serves, but this type of ideal reference population may not be accessible. For example, pediatric reference intervals are difficult to establish because of ethical concerns over performing unnecessary venipuncture on children, who cannot legally provide consent. Similarly, it may be difficult to recruit healthy elderly subjects to donate specimens for a reference interval study due to the high incidence of chronic disease in this group. For these and other reasons, ideal reference populations are often unavailable, and some compromises may be necessary in the selection of a suitable population on which to base reference intervals.

Manufacturers of in vitro diagnostic reagents are required to establish reference intervals as part of the application for approval by the Food and Drug Administration,[2] and this requirement is ordinarily met by collecting data from several sites at which the product is tested prior to market. The population used for these studies is usually larger and more diverse than the reference population available to any particular laboratory, but may not faithfully represent the patients in specific geographical and demographic areas. For this reason, laboratory practice standards included in the 1988 revision of the Clinical Laboratory Improvement Act (CLIA), originally passed by Congress in 1967, required clinical laboratories to verify that reference intervals were appropriate for their specific patient populations.[3]

Statistical methods can be applied to the task of establishing reference intervals, as well as their validation in individual laboratories.

Establishing Reference Intervals

Reference intervals customarily represent the central 95% of values obtained from the reference population. Consequently, 2.5% of "normal" individuals will exceed the reference range, and 2.5% will be below it. It is tempting to assume that normal values for clinical laboratory measurements conform to a Gaussian distribution, in which the central 95% of the area under the probability distribution curve corresponds to the population mean (μ) ± 1.96 standard deviations (usually rounded to 2 SD, or 2σ). However, this approach is often misguided, since the concentrations of various biochemicals in the body rarely follow a Gaussian distribution, due to physiological factors that influence the concentration in a unidirectional manner; intra-individual variations are not strictly random. Statistical approaches that are based on a predictable distribution of data, such as the Gaussian (or "Normal") distribution, are called "parametric," since they make certain assumptions about the data derived from the population. Non-parametric methods make no assumptions about how the data are distributed, and provide ways to analyze and compare data sets that have unknown or unpredictable distributions.

The Gaussian Distribution

Probability distributions are powerful statistical tools that allow predictions to be made about population data. A statistical distribution is a mathematical probability function that describes the relationship between the value of a particular measurement and the probability that randomly selected data in the population will have that value. The familiar Gaussian distribution is simply a mathematical probability function that expresses the relationship between the mean (μ) and standard deviation (σ) of a set of data, and the probability that a randomly selected data point will have a particular value, x. Gaussian distributions are characteristic of data that are influenced by multiple, random, independent errors in measurement. The first property that is noticeable when $P(x)$ (probability) is plotted against $(x\text{-}\mu)/\sigma$ (**Figure 1**), is the symmetry of the resulting bell-shaped curve; in a Gaussian distribution, $P(x\text{-}\mu) = P\text{-}(x\text{-}\mu)$. This property follows directly from the way in which the Gaussian probability function is derived, requiring that factors influencing individual measurements are random and independent. In a Gaussian distribution, the central 95% of data are bounded by the approximate limits $\mu \pm 2\sigma$, where μ is the population mean and σ is the population standard deviation. Therefore, if a reference range study for plasma glucose concentrations in healthy, non-diabetic individuals generated a mean of 91 mg/dL and a standard deviation of 8 mg/L, and a Gaussian distribution of the data was assumed, then the reference interval would be 91 ± 2(8), or 75 – 107 mg/dL.

Is it a valid assumption that plasma glucose in healthy individuals would be distributed in a Gaussian fashion? Probably not, because the factors that influence plasma glucose concentration are neither strictly random, nor entirely independent. Age and obesity, for example, are factors associated with impaired glucose tolerance even in non-diabetic individuals, and these factors do not have a random influence on plasma glucose (nor are they completely independent, since older patients are more likely to be overweight). Both of these factors *increase* glucose levels, so any distribution of plasma glucose concentrations in non-diabetic individuals would almost certainly be skewed toward higher values, rather than symmetrically distributed around the mean. Glucose is not an atypical example. The concentrations of most clinically relevant analytes in healthy individuals have distributions that are skewed toward higher or lower values, owing to physiological factors that have a strictly unidirectional influence on their concentration.

Log Transformation

In cases where the distribution of normal values is heavily skewed toward higher results, a plot of the log concentration vs. frequency may produce a curve that is more symmetrical and similar to a Gaussian distribution (**Figure 2**). If the resulting log-transformed distribution appears Gaussian, then some of the useful properties of Gaussian distributions, such as the $\pm 2\sigma = 95\%$ rule, may be applied. In 1972, Harris and DeMets[4] proposed log transformation as a means for generating a symmetrical distribution of reference values. It is important to remember, however, that whether or not data conform to a Gaussian distribution is determined by the randomness and independent nature of the influences that cause variation in the data points, and mathematical transformation of the data does not change those fundamental influences.

Non-Parametric Reference Ranges

If a distribution is not Gaussian, the central 95% of the data can be determined by ordering the array from the lowest to the highest values, and eliminating the highest 2.5% and lowest 2.5% of values; the remaining highest and lowest values delimit the reference interval. Non-parametric methods do not make any assumptions about the distribution of values in the data set, such as whether it is symmetric about the mean and whether the distribution is skewed toward higher or lower values. Although non-parametric determination of reference intervals is a simple and straightforward procedure that does not rely on any assumptions about the distribution, the method has some limitations.

One limitation is that non-parametric methods ignore any errors associated with individual measurements. Using the example of plasma glucose measured in 40 healthy volunteers, if the data are arranged from lowest to highest glucose concentrations, the non-parametric reference interval would be defined by the 2nd and 39th values in the ordered array, since 2.5% of 40 = 1.

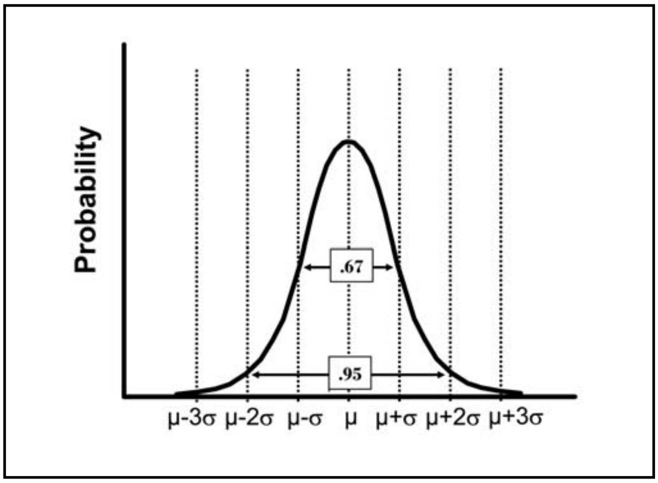

Figure 1_The Gaussian probability distribution curve (also called the "normal" or "bell-shaped" curve). The population mean (μ) is at the center of the symmetrical distribution, and 67% of the area under the curve falls within one standard deviation (σ) on either side of the mean. Approximately 95% of the area under the curve is between $\mu - 2\sigma$ and $\mu + 2\sigma$. Gaussian probability distributions predict the variability in measurements that are affected by multiple, independent, random errors. Laboratory quality control is a good example of a process to which a Gaussian distribution should apply. Inter-individual variations in clinical analytes do not ordinarily follow this distribution.

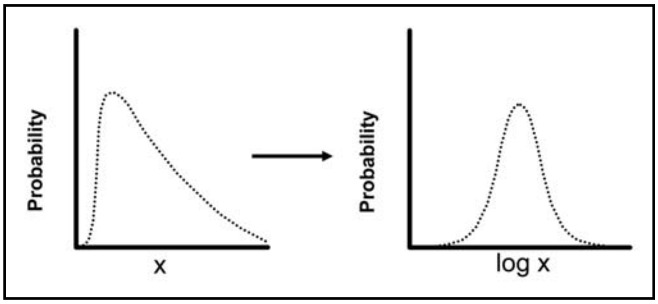

Figure 2_Log-transformation of data skewed toward higher values. In some cases, skewed data can be made more symmetrical by log transformation, usually for the purpose of applying Gaussian statistics to the data.

But there is some variance associated with each of those data points. The 2nd and 39th values in the ordered array will only be approximations, within the limits of the precision of the assay, of the 2.5th and 97.5th percentiles; the non-parametric statistical method does not account for those variations. In contrast, the Gaussian distribution takes into account all random influences in determining the upper and lower limits of the central 95% of values.

The variability associated with individual data points can be minimized, to a degree, if the dataset is very large. Therefore, in order to produce reliable limits for the 2.5th and 97.5th percentiles, non-parametric distributions require fairly large reference populations. The Clinical and Laboratory Standards Institute (CLSI; formerly the National Committee on Clinical Laboratory Standards, NCCLS) recommends that reference intervals be determined by a non-parametric method, with data from at least 120 appropriately selected subjects. The 3 highest and 3 lowest values are eliminated, and the 4th and 117th numbers in the ordered array define the reference interval.

Validation of Reference Ranges

Part of the data supplied to the Food and Drug Administration (FDA) in a 510(k) application for approval of an in vitro diagnostic method is a reference interval determined with the proposed method. These reference interval studies may be conducted in the hospital laboratories where the reagents are evaluated, and may use patient specimens or healthy volunteers. Manufacturer-determined reference intervals are typically based on a large number of specimens (often a thousand or more), and the proposed normal range is included in the product literature. Current CLIA guidelines require that laboratories using a manufacturer's reference interval—or, for that matter, any reference interval that is transferred from an external source—verify that it is appropriate for the population served by the laboratory. Laboratories must determine whether a reference interval based on data from the manufacturer's "healthy" population is the same as the reference interval for the population that the laboratory serves. Although it is possible to meet this requirement without gathering reference data from a local population, validation of a reference interval ordinarily involves collection of specimens from healthy volunteers, and comparison of the results to the proposed reference interval. Alternatively, a laboratory may establish its own reference interval by collecting 120 specimens, as recommended by CLSI, but this may be an impractical alternative for many laboratories.

If the concentrations of various analytes in healthy individuals followed Gaussian probability distributions, then the reference intervals could be compared by several parametric statistical methods. The Student's *t* test, for example, estimates the degree to which a small sample selected from a population predicts the properties of the entire population (specifically, the μ and σ). With regard to reference intervals, the question is whether the statistical characteristics (μ and σ) of a small sample of healthy individuals selected locally match the population statistics on which the manufacturer's (or other laboratory's) reference interval is based. Parametric methods provide ways to make those comparisons, based on the variability in the mean and standard deviation that is predicted when a subset is randomly selected from population data. But parametric methods assume that the data have a predictable distribution, and as mentioned before, this is not usually the case for laboratory tests.

Non-Parametric Methods for Comparing Data

Just as there are both parametric and non-parametric statistical methods for determining the reference interval, both approaches exist for comparing data sets, as well. Non-parametric methods can be applied to determine whether 2 data sets have essentially the same, or significantly different properties. In the case of validating a reference interval, the 2 data sets may be the manufacturer's data, used to determine the suggested reference interval, and a sample of healthy individuals recruited locally by the laboratory.

The Mann-Whitney Test

An example of a non-parametric statistical method to compare data sets is the *Mann-Whitney test*. In this method, the 2 data sets to be compared—$x_1, x_2 \ldots x_N$ and $y_1, y_2 \ldots y_N$—are ordered, together, from the lowest to highest values. The array might look something like:

$$x_1, y_1, x_2, x_3, y_2, x_4, y_3, y_4, y_5, x_5 \ldots etc.$$

For the Mann-Whitney test, the total number of *y* values that follow each *x* value are summed, and likewise for the *x* values that follow each *y*. If these sums, U_x and U_y, are similar, then the 2 samples appear to be equivalent. Large differences between U_x and U_y indicate that the 2 data sets are not equivalent. The Mann-Whitney test is also called the *U-test*, *Rank Sum test*, or *Wilcoxen's test*.

The Run Test

Another non-parametric approach to comparing data sets is the *Run test*. As with the Mann-Whitney test, data from both arrays are ordered from lowest to highest, and the numbers of "runs," or sequential data elements from one or the other array, are counted. Two data sets selected randomly from a common population will produce few runs, whereas a significant bias between the 2 data sets will be reflected in the magnitude and inequality when the 2 run sums are compared.

It may be helpful to think about the Mann-Whitney and Run tests as statistical methods not so much for determining whether 2 data sets have the same mean and standard deviation, but rather a reflection of the degree to which the 2 data sets have the same distribution of values, which for non-parametric distributions is the more important question.

The Monte Carlo Method

Monte Carlo simulations make use of random selection to generate a representative statistical distribution that can be applied to solve a quantitative statistical problem. Although random sampling, as a method to generate statistical distributions, had been used by mathematicians since the 19th Century, credit for refining (and naming) this technique is usually given to Stanislaw Ulam, a Polish born mathematician who worked for John von Neumann on the Manhattan Project during World War II, and his collaborator Nicholas Metropolis, who published their description of Monte Carlo simulations in 1949.[5] The Monte Carlo method is an elegant approach to validating reference intervals, and an application to this problem was described by Holmes and colleagues in 1994.[6]

In the Monte Carlo approach, a limited normal range study is performed, perhaps involving 20 healthy volunteers selected from the local population served by the laboratory. The mean and standard deviation is calculated based on the in-house study. Then, using the larger data set on which the manufacturer's reference interval is based, 20 individual data points are randomly selected and the mean and standard deviation of this random sample is calculated. This procedure is repeated many times using computer algorithms for randomly selecting data points and calculating the mean and standard deviation based on those data subsets. When a sufficient number of samples have been selected from the parent (or population) data set, then the variance associated with the mean and standard deviation of a randomly selected 20 data point subset can be calculated. If the results for the local sample are truly representative of the population on which the manufacturer's reference interval is based, then the mean and standard deviation of the in-house study sample will fall within limits predicted by the Monte Carlo simulation. In other words, the statistical properties—mean and standard deviation—of the local population will appear equivalent to a randomly selected subset of the larger population on which the manufacturer based its reference interval. The power of this method is that it is entirely non-parametric, but requires only a small set of in-house data.

CLSI-Recommended Methods for Validation of Reference Intervals

Guidelines are available for the validation of reference intervals from the CLSI document C28-A, which describes 3 methods for meeting the CLIA-specified requirement.[7] Some of these recommendations have a basis in statistical theory, whereas others do not. Berry and Westgard reviewed extensively the CLSI recommendations for reference interval validation on the Westgard QC Web site.[8]

Inspection method. The demographic and geographic factors associated with the reference population are examined to determine whether they are consistent with the population served by the laboratory. If there are no credible reasons to suspect that the population served by the laboratory differs from the reference population in any manner that would affect the predicted results of a particular test, then use of the reference range may be justified. The CLIA guidelines allow the medical director of a laboratory to make that assessment.

The inspection method is not a statistical approach, and transference of a reference interval from one laboratory to another should not be done without a firm basis on which to conclude that the reference populations are similar. This method should only be used when reference data from local volunteers are unavailable. This may be the case, for example, with age-specific reference ranges for pediatric populations.

Limited validation. In a limited validation study, approximately 20 reference samples are collected from healthy volunteers selected from the population served by the laboratory. If no more than 2 measurements fall outside the reference interval, the range is validated. If 3 or more reference specimens are outside of the reference range, 20 additional reference samples can be obtained, and if 3 or more of the second reference sample are out of the reference interval, the laboratory should consider establishing its own reference range.

The limited validation is based on the statistical prediction that 19 of 20 randomly-selected data points should fall within the central 95% of values in a population. This prediction is regardless of whether the reference interval was obtained by parametric or non-parametric methods. The probability that fewer than 3 out of 20 randomly selected data points will fall outside of the central 95% limits is more than 90%. The probability of randomly selecting 3 or more values outside of the central 95% of the array on 2 consecutive trials of 20 is only about 1%, so failure of the second trial would lead one to conclude that the populations are sufficiently different to warrant a local reference interval.

Extended validation. Sixty reference specimens are obtained from healthy volunteers within the laboratory's catchment area, and the reference interval for the local population is calculated. If the reference interval is calculated parametrically with the assumption that the population data have a Gaussian distribution (95% limits = $\mu \pm 2\sigma$), then a sample of 60 data points randomly selected from the population should produce essentially the same reference interval. This is because, as a general rule, samples of greater than 30 data points randomly selected from a Gaussian population will have statistical properties that are representative of the entire population (this is predicted by the Student's *t* distribution). In other words, the mean and standard deviation of a subset of 30 or more data points are very close to the mean and standard deviation of the population. The Student's *t* distribution takes into account deviations from Gaussian behavior when the number of sample data is fewer than 30.

Non-parametric statistical methods, such as those described above, also can be used to compare the locally generated reference interval with the manufacturer's proposed interval. In either the limited or extended validation methods, outliers may be removed from the dataset by application of the "Reed rule": If the difference between the extreme value and the next closest value in the array is *D*, and the range between the lowest and highest values in the entire array is *R,* Reed's rule is violated when the ratio *D/R* exceeds one-third, and data points that violate this criteria can be eliminated.

Summary

Statistical analysis is helpful for characterizing, and in some instances predicting, the behavior of data sets. Establishing and validating reference intervals are tasks to which statistical analysis can be applied, since the fundamental purpose of a reference interval is to predict the results of laboratory tests in healthy patients. The "central 95% of healthy individuals" that customarily defines reference intervals is a compromise between the sensitivity (ability to detect disease) and specificity (ability to rule out disease) of a laboratory test. Adopting this definition of reference intervals ensures at least 5% of results will be falsely positive, but allows for some overlap between the distributions of positive and negative ("normal") results in order to improve the clinical sensitivity of the test. Because the limits of the reference interval ultimately define the sensitivity and specificity of a laboratory test, it is very important to apply the appropriate statistical method when determining these limits.

The distributions of most clinically relevant analytes in blood, urine, or other body fluids, do not have mathematically predictable properties. As a result, parametric statistical methods, which are based on mathematical probability functions that assume a predictable distribution of data, are not ordinarily applicable to the determination of reference intervals. Non-parametric statistical methods, which are applicable to any distribution of data, are preferable for determining reference intervals, but have limitations of their own, including the large number of data points necessary for generating a valid range. CLSI recommends that non-parametric reference intervals are based on 120 specimens from healthy volunteers representing a broad demographic profile.

Many laboratories use manufacturer-specified reference intervals, since these are based on large data sets. However, CLIA requires clinical laboratories to verify that their reference ranges are appropriate for the patient population they serve. Non-parametric statistical methods exist for comparing data sets, and these can be applied to validation of reference intervals when reference data are obtained from the local healthy population. Monte Carlo simulation is another non-parametric method for validating reference intervals when a small sampling is obtained locally. The CLSI provides some guidance on validating reference intervals, including simple inspection, limited validation, and extended validation. LM

1. Fraser CG. Inherent biological variation and reference values. *Clin Chem Lab Med.* 2004;42:758-764.
2. FDA 510(k) requirements. Available at: http://www.fda.gov/cdrh/manual/ivdmanul.html. Accessed on March 10, 2006.
3. CLIA '88. Available at: http://www.fda.gov/cdrh/CLIA. Accessed on March 10, 2006.
4. Harris EK, DeMets DL. Estimation of normal ranges and cumulative proportions by transforming observed distributions to gaussian form. *Clin Chem*. 1972;18:605-612.
5. Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc*. 1949;44:335-341.
6. Holmes EW, Kahn SE, Molnar PA, et al. Verification of reference ranges by using a Monte Carlo sampling technique. *Clin Chem*. 1994;40:2216-2222.
7. CLSI Document C28-A: How to define, determine, and utilize reference intervals in the clinical laboratory; Approved guideline. 1995.
8. Barry PL, Westgard JO. Method validation: Reference interval transference. Available at: http://www.westgard.com/lesson30.htm. Accessed on March 10, 2006.