Chapter 2
LABORATORY METHODS

LABORATORY MEASUREMENTS

The majority of the studies performed in the clinical laboratory consist of measurements of the concentrations of blood cells and of chemical substances in body fluids (Table 2.1). The chemical substances that are studied include gases, electrolytes, metabolic intermediates, waste products, tissue proteins, plasma proteins, hormones, micronutrients, drugs, and toxins. Most of what is discussed in this chapter applies most directly to such concentration measurements. However, with some modifications, the concepts can also be applied other kinds of laboratory measurements.

THE PROCESS OF MEASUREMENT

The process of laboratory measurement proceeds in four steps: sample preparation, analyte separation, analytical signal production and detection, and calculation of results. Sample preparation consists of the operations that must be performed on a specimen to yield the test material that constitutes the sample. It is the sample, or a portion of the sample, that is introduced into the measurement system and on which the measurement is carried out. In the determination of the plasma concentration of a chemical substance such as creatinine, sample preparation consists of the separation of the plasma (or serum if the specimen is clotted blood) from the blood cells by centrifugation of the specimen.

There may be an analyte separation step that represents a more-or-less specific isolation of the analyte of interest from the other chemical substances in the sample. For instance, if creatinine is to be measured using the Jaffé reaction, it can be absorbed to porous aluminum silicate clay or a cation exchange resin to separate it from other plasma substances that react with the signal generating reagent.

Analytical signals are generated and detected in a wide variety of ways. In automated blood cell counting, for instance, the signal consists of a voltage pulse arising from a change in electrical impedance across an aperture as a cell passes through the aperture. The signal is detected by a voltmeter. One way to produce an analytical signal for the measurement of chemical substances is by the formation of a chemical species that absorbs light. In the Jaffé reaction for creatinine, picrate reacts with creatinine under alkaline conditions to form a strongly light-absorbing red-orange compound, the

Table 2.1
Concentrations of selected blood analytes

Blood cells		
cells/L	10^{13}	
		red cells
	10^{12}	
	10^{11}	platelets
		reticulocytes
	10^{10}	
		neutrophils
	10^{9}	lymphocytes
	10^{8}	eosinophils
Chemical substances		
mol/L	10^{0}	
	10^{-1}	sodium, chloride
	10^{-2}	
		potassium, glucose, urea
	10^{-3}	calcium, carbon dioxide
		albumin
	10^{-4}	uric acid
	10^{-5}	bilirubin, iron, haptoglobin
		fibrinogen
	10^{-6}	
		cortisol
	10^{-7}	thyroxine
		hydrogen ion
	10^{-8}	
	10^{-9}	testosterone, factor VIII
		cobalamin
	10^{-10}	ferritin
	10^{-11}	
		parathyroid hormone
	10^{-12}	

Janovski complex. The signal consists of the reduction in the amount of light passing through the reaction solution. The signal is detected by a spectrophotometer.

Calculation of results

In order to arrive at a study result, the magnitude of the signal generated by a sample must be converted to a concentration value. This can be done in either of two ways. If the measurement system has been shown to have signal generating properties that closely match the theoretical ideal, the theoretical relationship between analyte concentration and signal magnitude can be used to calculate the result. For example, the theoretical relationship between analyte concentration and light absorbance is embodied in the Beer-Lambert law,

$$\textit{analyte concentration} = \frac{\textit{absorbance}}{\textit{analyte absorptivity} \cdot \textit{light path}}$$

For an ideal system, dividing the observed absorbance value by the known values for the absorptivity of the analyte and the length of the light path in the detector yields the analyte concentration.

In practice, the behavior of measurement systems is rarely ideal, so the use of theoretical relationships is not a satisfactory way to calculate results. Instead, results are calculated using the empirical relationship between analyte concentration and signal magnitude as established by the measurement of signals produced by a set of test materials with known analyte concentrations. These test materials are called calibrators (formerly, standards) and the relationship between analyte concentration and signal magnitude is called a calibration curve. A hypothetical linear calibration curve is shown in the upper graph of Figure 2.1

To convert the signal generated by a test sample to a concentration value, the calibration curve is used in reverse. That is, rather than finding the y (signal) value on the curve for a known x (concentration) value, an x value on the curve is found that corresponds to the known y value. For a signal of 100 units, for instance, the corresponding point on the calibration curve has a concentration of 10 units. The measurement result is therefore, 10 concentration units. Because it is unconventional to reverse the roles of the x- and y-axes, results are calculated using what is called a measurement curve which is identical to the calibration curve except that the x- and y-axes are switched (the lower graph in Figure 2.1). Note that inversion of the equation describing the calibration curve yields the equation that defines the measurement curve.

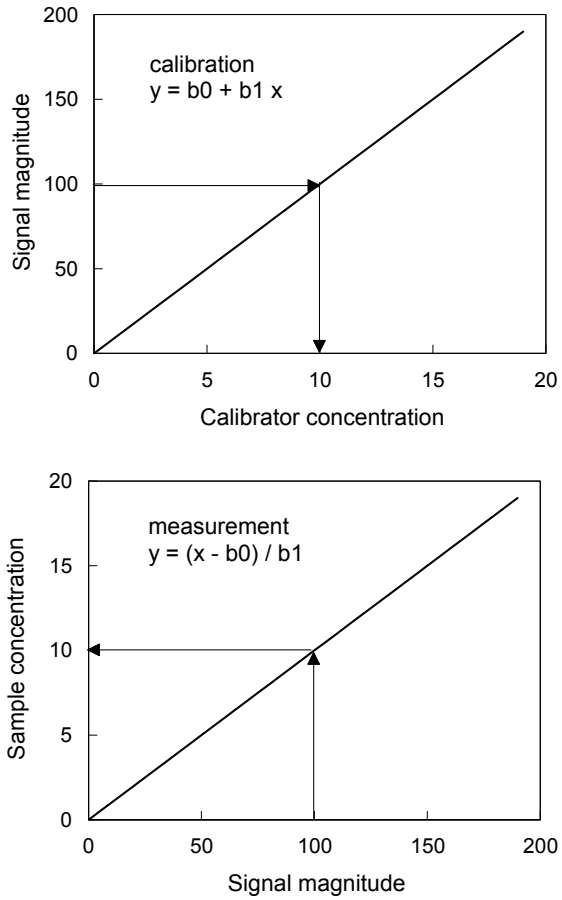


Figure 2.1 Hypothetical calibration and measurement curves (solid lines). The graphical technique for the calculation of the result given a signal of 100 units is depicted.

Calibration and measurement curves are usually constructed each time that a batch of measurements is made. Calibrators are run along with the test samples and the parameters of the equations defining the curves are estimated from the observed calibrator and signal magnitude pairs. The number and spacing of the calibrators should be chosen with the intent of providing for highly reliable estimation of the equation parameters. According to the statistical theory of optimal design, there exists a unique set of calibrator concentrations that yields the most precise estimates of the equation parameters (Fedorov 1972, Steinberg and Hunter 1984). In general, the number of separate concentrations that should be evaluated equals the number of parameters in the equation. In the case of a linear calibration curve there are two

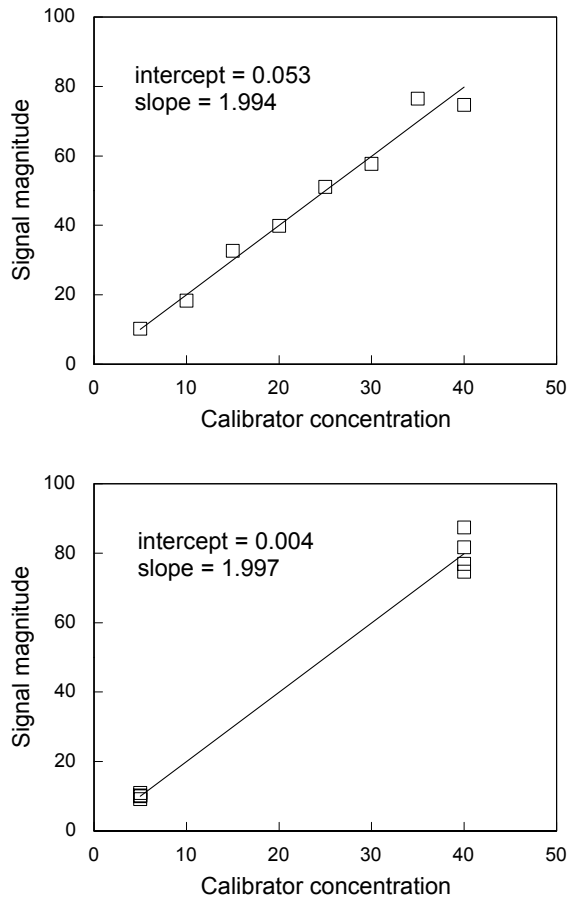


Figure 2.2 Calibration curves as obtained by two different calibrator spacing schemes. The even spacing scheme is shown in the top graph and the end-of-range spacing scheme is shown in the bottom graph.

parameters, the slope and the intercept, so two concentrations should be evaluated. The optimal concentrations are the concentrations at each end of the range of measurement. Figure 2.2 illustrates a simulated application of this optimal design. The hypothetical linear calibration relationship has an intercept of zero and a slope of two and proportional measurement variability. Using a scheme in which the calibrators are evenly spaced over the clinical range results in the parameter estimates shown in the upper graph. Using the optimal scheme (lower graph) with the same number of calibrators gives parameter estimates that are closer to the true (simulation) values. An added benefit of the end-of-range calibrator spacing scheme is that the inconstancy of the measurement variability is much more clearly demonstrated than with the even spacing scheme. Optimal spacing schemes can also be devised for more complex equations such as those for immunoassays (Bezeau and Endrenyi 1986).

Ordinary linear regression is the technique most frequently used for the estimation of the parameter values of linear calibration curves. This technique has an underlying assumption: the variability in the measurement of the dependent variable has a constant variance (Berry 1993). If the variability in the measurement of the analytical signal is not constant over the range of calibrator concentrations employed, weighted linear regression analysis, which adjusts for the inconstancy in the variability of measurement, should be used instead.

Nonlinear regression analysis is the most accurate and precise technique of parameter estimation for nonlinear curves (Motulsky and Ransnas 1987). It is the technique that is used to estimate the parameters of the sigmoidal calibration curves found in immunoassay systems

THE METHOD OF MEASUREMENT

Most measurements can be performed in a number of ways that vary in the nature of the sample, the technique of analyte separation, the means of producing and detecting the analytical signal, and the technique for calculating results. A specific way of performing a measurement is referred to as a method of measurement or, informally, as a laboratory method or, more simply still, as a method. Laboratory methods are the basic units of laboratory practice.

METHOD QUALITY

The quality of a laboratory method can be defined as the ability of the method to satisfy the clinical needs served by measurements made with the method. There is always a need for trueness and precision.

Trueness of measurement

Trueness is defined as the closeness of agreement between the true value of an analyte and the average result value obtained from a large number of replicate measurements (Stöckl 1996, Dybkaer 1995). It is the term that is currently applied to the concept that used to be called accuracy. Accuracy is now used to denote the broader concept of closeness of agreement between the true value of an analyte and a single measurement result. As such, accuracy reflects both trueness and precision of measurement. Trueness is measured on a ordinal scale of the sort:

poor/average/excellent. The inverse of trueness, which is called bias (or systemic error of measurement), is a quantitative measure. It is the difference between the true value of an analyte and the average result value. Because it is useful to have a quantitative measure of quality, bias is the measure that is applied when appraising the trueness of a measurement method.

Bias arises when the calibration process does not reflect the test measurement process perfectly. Causes of bias include matrix effects, calibrator effects, and treatment effects (Strike 1996).

Matrix effects arise from the differences that exist between the complex biological matrix found in test samples and the artificial matrix of the calibrators. Physical matrix effects are caused by physical properties, such as sample viscosity, that result in test samples being processed differently than calibrators by the measurement instrument. Nonspecific chemical matrix effects, called interference effects, are caused by substances in the test sample that, while not generating a signal themselves, affect the magnitude of the signal generated by the analyte being measured (Kroll and Elin 1994). Specific chemical matrix effects are referred to as cross-reaction effects. They are caused by substances in the test sample that generate a signal identical to that of the analyte of interest. Considerable effort is devoted to the evaluation of cross-reactions during the development of laboratory methods and various techniques may be employed to improve the specificity of the method by reducing or eliminating the cross reacting substances. Separation of the analyte from cross-reacting substances, discussed earlier as a frequent step in the measurement process, is one means of increasing method specificity. The selective adsorption of creatinine to fuller's earth was mentioned as an example. Some other separation techniques are listed in Table 2.2. One of these techniques, liquid chromatography, is even more successful than fuller's earth in specifically isolating creatinine from substances that cross-react in the Jaffé reaction. Method specificity can also be improved in the analytical signal generation and detection step of analyte measurement. One approach taken at this step is to increase the selectivity of the signal generating reaction so that only the specific analyte participates in the reaction. One way to do this is to use analyte-specific enzymes to catalyze a reaction that leads to the production of the signal. A number of enzymatic methods are available for the measurement of creatinine. In the most popular method, creatinine is hydrolyzed to creatine by the highly specific enzyme, creatinine amidohydrolase. Creatine formation is coupled, through a series of specific enzymatic reactions, to the production of a light-absorbing species that provides the analytical signal. The specificity inherent in enzyme reactions can also be taken advantage of when an enzyme is itself the analyte of interest. In that case, a reagent that is a substrate of the enzyme undergoes a catalytic conversion to a product that is coupled to the production of the analytical signal. For example, creatine kinase concentrations are determined by measuring the rate of formation of ATP and creatine from the substrates ADP and creatine phosphate. Note that, in methods of this sort, enzyme concentrations are measured and reported in terms of enzymatic activity rather than substance concentration. Another approach for improving method specificity in the analytical signal generation and detection step is to increase the selectivity of signal detection so that the only the signal generated by the analyte is detected. One way this is done is by the so-called kinetic technique which depends upon a differential rate of signal production for the analyte and cross-reacting substances. In the Jaffé reaction, the cross-reacting substances tend to react with picrate slowly compared to creatinine so the initial rate of Janovski complex formation is due largely to creatinine. By measuring the initial rate, the signal from creatinine can be segregated from the signals arising from the cross-reacting substances.

Bias due to calibrator effects arises from differences between the analyte used in the calibrators and the analyte as found in patients. One way this happens is when a class of chemical species represents the analyte of interest but the calibration material is based on a single species within the class. The measurement of total protein in the urine is a good example of this situation. Another cause of calibration effects is the use of non-human or altered human analyte in calibrators. Analytes used in calibrators must be available in quantity and they must be stable. It is often simply impossible to procure from human sources adequate amounts of trace analytes, such as hormones. It is similarly impossible to preserve in unaltered form fragile analytes such as blood cells.

Bias can also be caused by treatment effects. These effects arise when test specimens and calibrators are not treated in an identical fashion when

Table 2.2
Selected Separation Techniques for Improving Method Specificity

Technique	Principle and examples
Membrane permeability	analyte-permeable membrane separates sample from site of signal generation only specific analyte passes through membrane O_2 and CO_2 electrodes, ion-selective electrodes, ultrafiltration, equilibrium dialysis
Electrophoresis	chemicals in buffer migrate through support medium in electric field chemicals have different migration rates due to different charges and different degrees of interaction with stationary phase specific analyte migrates to characteristic location cellulose acetate electrophoresis, agarose gel electrophoresis (named for support medium)
Chromatography	mobile phase (gas or liquid) passes through a column with a stationary phase bound to support medium chemicals in mobile phase have different degrees of interaction with stationary phase which leads to different migration rates through column specific analyte elutes from column at characteristic retention time gas chromatography, liquid chromatography (named for mobile phase)
Antibody binding	analyte-specific antibodies are bound to support medium only specific analyte binds to antibodies and is retained on support medium following wash heterogeneous radio-, enzyme, and fluorescence immunoassays

preparing analytical samples. For example, prior to measuring the concentration of some protein-bound and intracellular analytes, it is necessary to release the analyte into solution. To accomplish this, a release reaction will be performed on the test specimens. The same reaction may not be performed on the calibrators because the analyte in calibrators is already in solution.

Precision of measurement

Precision is defined as the closeness of agreement among the result values obtained from a large number of replicate measurements (Dybkaer 1995). Precision is measured on an ordinal scale but its inverse, imprecision, is a quantitative measure. Imprecision (or random error of measurement) is the dispersion of results for a large number of replicate measurements. It is usually expressed in terms of standard deviations.

Imprecision arises from multiple sources which can be categorized according to the following scheme (Dybkaer 1995): (1) those that arise during a single batch of measurements, (2) those that arise over the course of the performance of multiple batches of measurement, and (3) those that arise when several laboratories contribute to the production of results.

Imprecision arising during a single batch of measurements is called within-run, or within-batch imprecision and is measured in terms of the within-run standard deviation. Because it quantifies method precision in the setting of the minimum number of sources of measurement variability, within-run imprecision is the minimum precision attainable by the method. The causes of within-run imprecision include volumetric errors, instrumental fluctuations, variability in the efficiency of the separation step, and vagaries in the rate and completeness of the signal generating step.

Imprecision arising from the performance of multiple batches of measurement is called between-run, or between-batch, imprecision and is measured in terms of the between-run standard deviation. The causes of between-run imprecision include recalibration of the measurement system, different calibrators and reagents, different operators, and time

dependent phenomena such as drift in the performance characteristics of the measurement instrument.

Within-laboratory imprecision is the total imprecision in the measurement of an analyte in a single laboratory. It reflects the variability arising within and between runs. The variances of these two sources add together to give the total variance,

$$var_{within\text{-}laboratory} = var_{within-run} + var_{between-run}$$

where *var* is variance. In terms of the usual measure of imprecision, standard deviations,

$$SD_{within\text{-}laboratory} = \sqrt{SD^2_{within-run} + SD^2_{between-run}}$$

where *SD* is standard deviation.

The imprecision that arises when several laboratories contribute to the production of results is called between-laboratory imprecision. It is caused by inter-laboratory variation in calibrators, calibration spacing scheme, choice of calibration function, and technique for estimation of calibration curve parameters. Other causes include differences in operating conditions, differences in operator skill, and differences in the measurement system.

Resolving power. The precision of a method determines how good the method is at distinguishing differences in analyte concentration. This property, referred to as the resolving power of a method, is a useful alternative measure of method precision, especially when small changes in analyte concentration must be discerned and when trace concentrations of analyte must be detected (Sadler *et al.* 1992, Gautschi *et al.* 1993). The resolving power of a method is what is often referred to as the analytical sensitivity of the method (Ekins and Edwards 1997). Resolving power is a less confusing term, however, because analytic sensitivity is also taken to mean the slope of the calibration curve (Pardue 1997).

The usual way in which resolving power is expressed is as the minimum distinguishable difference in concentration, D_{min}. This parameter can be defined for within-run differences or for between-run differences. For between-run differences (Sadler *et al.* 1992),

$$D_{min} = z_c \sqrt{2}\, SD_{within\text{-}laboratory}$$

where z_c is the confidence coefficient as found with the standard normal distribution; z_c equals 1.645 for a 95% confidence level. This formula assumes that method precision is essentially constant over intervals of analyte concentration equal in length to D_{min}. The detection limit of a method, which is the smallest analyte concentration that can reliably be distinguished from zero, is a special case of D_{min}.

Hierarchy of method quality

The ideal of laboratory practice is to implement methods of the highest quality. Unfortunately, of the methods available for the measurement of a particular analyte, those of the very highest quality are always too expensive and too impractical for most clinical laboratories. These methods, which are called definitive methods, are used to validate the accuracy of the methods at the next level of quality, called reference methods. Reference methods, which have only negligible inaccuracy compared to definitive methods, are generally less costly than definitive methods but they are still impractical for routine use. They are used to validate the accuracy of the affordable and practical methods of lower quality that are actually implemented in the clinical laboratory. These methods are called field methods. This hierarchic chain of validation of the accuracy of laboratory methods represents one of the two elements of the system of accuracy transfer that is used to assure the quality of field methods. The other element of the system is a hierarchy of calibrators. In this hierarchy, field methods are calibrated with secondary reference materials, these being calibrators whose values have been established using a reference method. Reference methods, in turn, are calibrated with primary reference materials which are calibrators whose values have been certified by competent authority through the use of a definitive method.

Analytical quality goals

It is recognized that field methods cannot provide reference method-level analytical quality given the constraints of affordability and practicality within which the methods must operate. However, it is necessary that the methods achieve a minimum level of quality—one that allows them to be of use clinically. It is therefore useful to define a desirable level of quality that can be used by both method developers and laboratorians as a benchmark for field method performance.

A number of different approaches can be used to define desirable analytical quality goals (Stöckl *et al.* 1995). These approaches include defining goals in keeping with the current "state of the art" in high-quality laboratories, having experts define the goals, and basing goals on the quality expectations of

clinicians. Additionally, quality goals can be derived from a consideration of the biologic variability of the analyte being measured. For instance, as discussed in Chapter 1, the extent of the variability in study results with repeated testing of an individual is determined by the within-individual biologic variability of the analyte and the within-laboratory analytical variability,

$$SD_{within\text{-}individual} = \sqrt{SD^2_{within-individual,biologic} + SD^2_{within-laboratory}}$$

The fractional increase in the total within-individual variability attributable to analytic variability is, therefore,

$$\sqrt{1 + \left(\frac{SD_{within-laboratory}}{SD_{within-individual,biologic}} \right)^2} - 1$$

To keep the contribution of the analytical component at a reasonable level, say 10 percent, the ratio of the within-laboratory variability to the within-individual biologic variability must be less than 0.459. Rounding up to 0.5 (for which the fractional increase in within-individual variability is 11.8 percent) yields the quality goal for repeated testing (Cotlove *et al.* 1970, Harris 1979, Fraser *et al.* 1997),

$$SD_{within\text{-}laboratory} < 0.5\ SD_{intra\text{-}individual,biologic}$$

This rule can also be expressed as

$$CV_{within\text{-}laboratory} < 0.5\ CV_{intra\text{-}individual,biologic}$$

which is particularly useful if within-individual biologic variability is proportional to analyte concentration. Then both coefficients of variation will be constant.

The reference interval for an analyte depends upon the median analyte value and the total (intra- and inter-individual) biologic variability of the analyte. In also depends upon the bias in the measurement of the analyte and the within-laboratory analytical variability. When calculated based on the assumption of a normal frequency distribution,

$$reference\ interval = median\ value + bias \pm 1.96\ SD_{total}$$

where

$$SD_{total} = \sqrt{SD^2_{biologic} + SD^2_{within-laboratory}}$$

The presence of bias results in a displacement of the measured reference interval from the true reference interval which is,

$$median\ value \pm 1.96\ SD_{biologic}$$

As a result, at one side of the reference interval, individuals who fall inside the measured reference interval fall outside of the true reference interval. At the other side of the reference interval, individuals who are outside of the measured reference interval are inside the true reference interval. The fraction of the population misclassified in this way should be kept to an acceptable level. If a 5 percent misclassification rate is used as the standard, in the absence of analytical imprecision, the ratio of the bias to the total biologic variability should be kept less than 0.315. Rounding down to 0.25 (for which the misclassification rate equals 3.7 percent) yields the quality goal for reference intervals (Gowans *et al.* 1988 and 1989, Fraser *et al.* 1997),

$$bias < 0.25\ SD_{biologic}$$

which can also be expressed as

$$relative\ bias < 0.25\ CV_{biologic}$$

Analytic imprecision widens the reference interval and thereby also results in misclassification. Using 3.7 percent misclassification as the standard, in the absence of bias, the ratio of the within-laboratory variability to the total biologic variability should be kept less than 0.56. This is another quality goal for reference intervals (Fraser *et al.* 1997),

$$SD_{within\text{-}laboratory} < 0.56\ SD_{biologic}$$

which can also be expressed as

$$CV_{within\text{-}laboratory} < 0.56\ CV_{biologic}$$

Of course, the misclassification rate should also be within desirable limits when both bias and imprecision are present. Figure 2.3 shows the paired values for relative bias and imprecision that satisfy the 3.7 percent misclassification standard.

The application of these quality goals can be illustrated by using them to define the desirable analytical quality of a field method for plasma creatinine concentration. At a concentration of 100 μmol/L, the average $SD_{within\text{-}individual,biologic}$ is 4.3 μmol/L ($CV_{within\text{-}individual,biologic}$, 4.3%) and the average $SD_{biologic}$ is 11.3 μmol/L ($CV_{biologic}$, 11.3%) (Sebastián-Gámbaro *et al.* 1997). The quality goal for precision based on the rule for repeated testing is an $SD_{within\text{-}laboratory}$ of less than 2.15 μmol/L

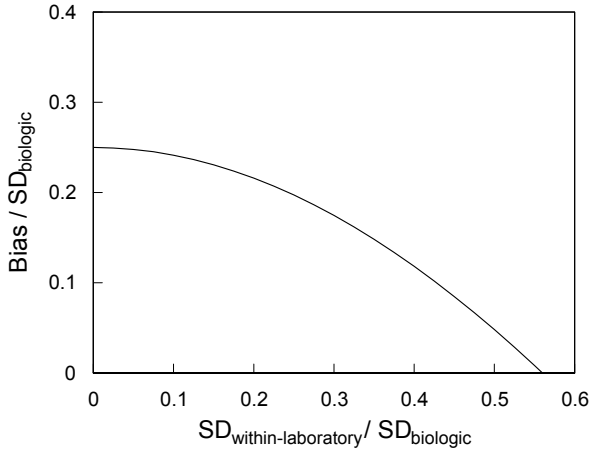


Figure 2.3 Analytical quality goals for method imprecision and method bias based on a consideration of patient classification using the reference interval for an analyte. Imprecision and bias are expressed relative to total biologic variability.

($CV_{within\text{-}laboratory}$ of less than 2.15%). At this level of imprecision, the ratio of $SD_{within\text{-}laboratory}$ to $SD_{biologic}$ is 0.19. So, using Figure 2.3, the quality goal for trueness based on the rule for reference intervals is a relative bias of less than 0.215 which corresponds to an absolute bias of less than 2.43 μmol/L.

Table 2.3
Components of a Written Measurement Procedure

Introductory material

- Title
- Table of contents
- Introduction
- Scope
- Warning and safety precautions
- Definitions
- Symbols and abbreviations
- References
- Dates

Method description

- Sampling and specimen handling
- Principle of measurement
- Reagents
- Apparatus
- Preparation of measurement system
- Use of measurement system
- Modifications of the usual procedure for special cases
- Calculation of results

Quality assurance program

Analytical performance description

- Analytical quality evaluation findings
- Method comparison findings

MAINTAINING QUALITY

Once a laboratory method has been implemented in the clinical laboratory, its quality is maintained by strict adherence to the approved procedure for performing the method, by the maintenance of high technical skill among the method operators, by the regular maintenance of the instruments utilized in the method, and by a rigorous quality assurance program.

Written measurement procedure

The laboratory document that contains the description of the steps in the performance of a method is called the written measurement procedure (Dybkaer 1997). In addition to describing the method, this document includes introductory material, a description of the quality assurance program for the method, and a summary report of the quality evaluation of the method (Table 2.3).

The introductory material identifies the method, summarizes the clinical rationale for the measurement of the analyte and the laboratory rationale for the choice of method, specifies the type of specimen, and stipulates safety precautions in the use of the method. The introductory material also provides lexicographic support for an unambiguous reading of the method description and lists of cited and recommended references. The procedure is periodically reviewed and is abridged as necessary. The dates of review and the dates and details of changes in the method are recorded and filed with the introductory material.

The method description is thorough and detailed. It includes sections devoted to specimen collection and handling, reagents and equipment, preparation and performance of the method, calculation of results, and quality assurance. Aspects of specimen collection to be considered include any special preparation of the patient for the taking of a specimen and the identification of the collection device and the specimen container. The handling of the specimen is detailed as regards anaerobic conditions, temperature, the allowable time prior to processing, the method of processing, and the conditions of storage of the processed specimen. The list of reagents stipulates the identity and source of each reagent and gives instructions for its storage, handling, and disposal. The preparation of stock and working solutions is described and the

shelf-lives of the solutions are indicated. The instruments and auxiliary equipment called for by the method are identified and their operation and maintenance are covered, often by reference to the manufacturer's manual. The procedure for readying the equipment and the samples (natural samples, blanks, calibrators, and controls) is indicated. The steps involved in the performance of the method are presented in a sequential fashion; included are the steps leading to the standby condition, if there is one, and the steps for closing down the method. If the operating steps are different in certain circumstances, the modifications of the method and the circumstances for their application are described. The mathematical methods for deriving the calibration function and the measurement function are described and the algorithm for the calculation of sample results is given. The computer software to be used to perform these calculations is stated.

Quality assurance program

In the broadest possible sense, quality assurance is concerned with the reliability of the patient data generated by the laboratory. It therefore encompasses the procedures used to recognize, quantify, and control the sources of measurement variability that arise within the laboratory between the receipt of a specimen and the posting of the study results (Büttner *et al.* 1980a). In its more common usage, quality assurance refers to the control of the precision and trueness of laboratory methods. It is in this more narrow sense of quality control that quality assurance will be discussed here.

Internal quality control

Internal quality control refers to the procedures for quality monitoring, intervention, and remediation undertaken in a single laboratory (Büttner *et al.* 1983a, Nix *et al.* 1987, Petersen *et al.* 1996). The unit of control is typically the set of samples that constitute one batch of the method. As mentioned previously, each batch includes a set of calibrators for the purpose of constructing a calibration curve for that batch. This is done to reduce the variability that results from between-run calibration variation in the method. As this is a quality maintenance goal, batchwise calibration is properly considered one of the elements of internal quality control.

Also included in each batch of samples is a set of control samples for which the measurement result frequency distributions are known. Using statistical tests called control rules to compare the current results for the control samples with their known frequency distributions, the trueness and precision of the method can be monitored.

Control samples are derived from control material rather than individual patient specimens but are otherwise handled in a fashion identical to test samples. Indeed, valid internal quality control depends upon the identical treatment of the control samples and the test samples. Control material must come from a large, homogeneous, and stable pool of material (Büttner *et al.* 1980c). The composition of control material should be as similar as possible to the composition of the test material used in the method being controlled. This requirement is best satisfied by control material made from human products but such material is generally biohazardous. Instead, the most widely used type of control material is commercially manufactured artificial material which is contrived to simulate the corresponding test material.

Using the techniques for the assessment of method quality discussed later in this chapter, the mean analyte concentration of the control material is established as is the within-laboratory imprecision in the measurement of the concentration. For quality control purposes, method trueness is then evaluated in terms of this mean and method precision is evaluated in terms of this $SD_{within\text{-}laboratory}$. This is done each time a new lot of control material is introduced into use in the laboratory.

Control rules. The logic of control rules is best appreciated by considering the effects that a decline in method quality has upon the frequency distribution of control sample results.

Degradation in the quality of a method may be characterized by an increase in method bias, by an increase in method imprecision, or by both. If there is an increase in method bias, the control sample results will tend to be displaced from the established mean value for the control material. This means that there will be an increased probability for the control sample results to be in the region of one of the tails of the established distribution. For instance, as shown in the graph on the left in Figure 2.4, if the current bias is equal to 1 $SD_{within\text{-}laboratory}$, 15.9 percent of the control sample results will be larger than the established mean plus 2 $SD_{within\text{-}laboratory}$. For the established distribution, only 2.3 percent of the control sample results would be that far above the mean. On the other tail of the distribution, the bias will

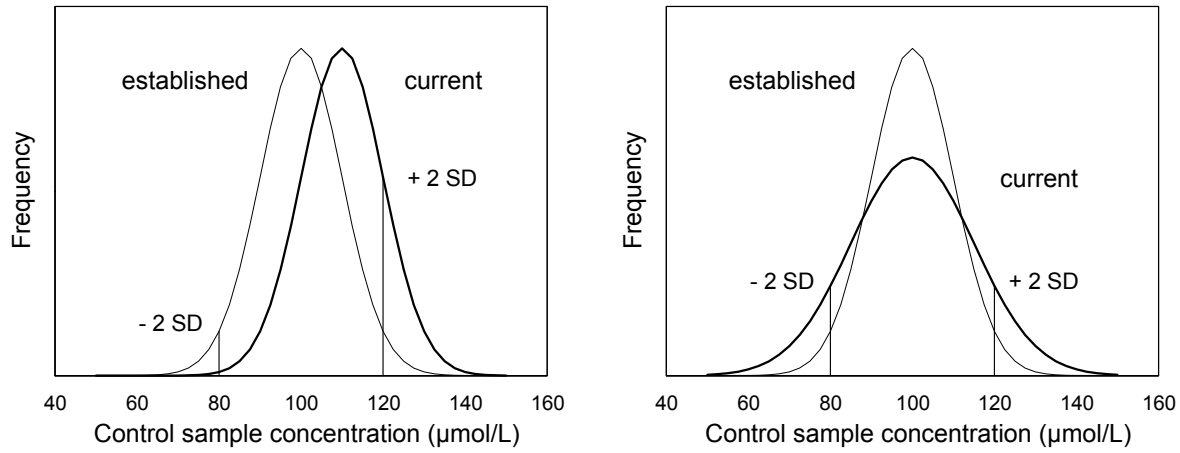


Figure 2.4 Frequency distributions of control sample results for a hypothetical laboratory method. The established distribution has a mean of 100 µmol/L and an $SD_{within\text{-}laboratory}$ of 10 µmol/L. It is shown as a light curve in both graphs. The current distributions are shown as dark curves. In the graph on the left, the method currently has a bias of 10 µmol/L. In the graph on the right, the $SD_{within\text{-}laboratory}$ of the method is currently 15 µmol/L.

result in only 0.1 percent rather than 2.3 percent of the control sample results having values smaller than the established mean minus 2 $SD_{within\text{-}laboratory}$. The net effect of the bias is an overall increase in the percentage of control sample results outside of the central region of the established distribution, 16.0 percent versus 4.6 percent. Identical percentages apply when the bias is negative.

If there is an increase in method imprecision, there will be an increased probability for the control sample results to be in the region of the tails of the established distribution. An example of this is shown in the graph on the right in Figure 2.4. For a 1.5-fold increase in imprecision, 9.1 percent of the control sample results will be more than 2 $SD_{within\text{-}laboratory}$ larger than the established mean and 9.1 percent of the control sample results will be more than 2 $SD_{within\text{-}laboratory}$ smaller than the established mean. Hence, the net effect of the increase in imprecision is an overall increase from 4.6 to 18.2 in the percentage of control sample results outside of the central region of the established distribution.

In general, there is an increased probability for control sample results to be more than 2 $SD_{within\text{-}laboratory}$ larger or smaller than the established mean if there is currently a bias in the method or an increase in the imprecision of the method. Therefore, a control sample result of this magnitude can be taken as a indication of a reduction in method quality. This is the statistical basis of the 1_{2s} control rule: a batch of measurements should not be considered to be in-control if one control result exceeds the mean plus or minus 2 $SD_{within\text{-}laboratory}$.

The performance of this control rule, or any control rule, is characterized by the relationship between the probability of rejecting a batch using the rule and the magnitude of the current change in quality in the method. This graphical presentation of this relationship is called the operating characteristic curve. Figure 2.5 shows the operating characteristic curves for the 1_{2s} control rule when using one to five control samples per batch. These curves represent the operating characteristics of the 1_{2s} control rule for detecting current bias; operating characteristic curves can also be drawn for the detection of increased imprecision and a 3-dimensional surface can be used to present the operating characteristics for the simultaneous detection of bias and increased imprecision. A bias of zero means that the quality of the method is unchanged so a rejection of a batch at this value represents a false rejection. Thus, the y-intercepts of the curves equal the rates of false rejection when using the indicated number of control samples.

In the quality control of a particular method, the control rule that should be used is the one that most closely achieves the clinical quality control goals for performance and false-rejection rate while minimizing the number of control samples run per batch. Consider, for instance, a method for which the false-rejection rate goal is 5 percent and the performance goal is 90 percent rejection of batches for which the current method bias is equal to or greater than 2.5 $SD_{within\text{-}laboratory}$. Using the 1_{2s} control rule, two control samples must be run per batch to achieve the performance goal at the stipulated level of bias.

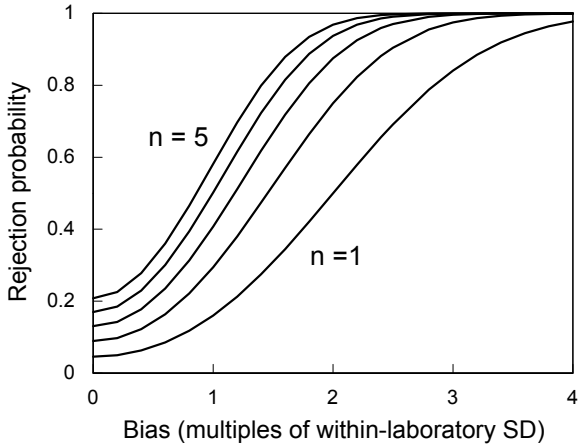


Figure 2.5 Operating characteristic curves for the detection of current method bias using the 1_{2s} control rule. Curves are shown for one to five control samples per batch.

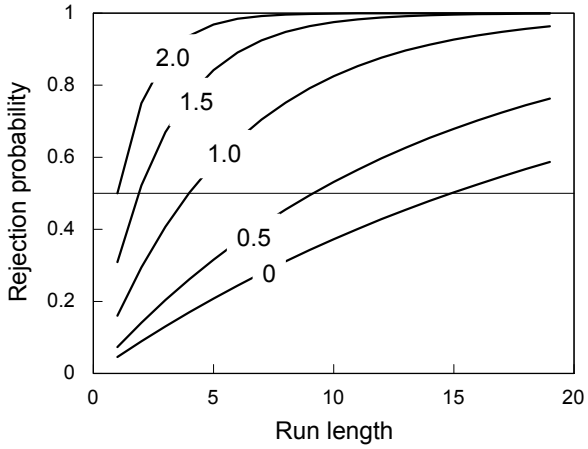


Figure 2.6 Cumulative run length distributions for the detection of persistent method bias using the 1_{2s} control rule. Curves are shown for method bias in half multiples of $SD_{within\text{-}laboratory}$.

However, with two control samples per batch, the false-rejection rate is too high, 8.9 percent. Using the 1_{3s} control rule (which states that a batch of measurements should not be considered to be in-control if one control result exceeds the mean plus or minus 3 $SD_{within\text{-}laboratory}$), an acceptable false-rejection rate, 1.6 percent, can be achieved at the performance goal, but at the expense of requiring six control samples per batch. Fewer control samples per batch can be used while still achieving the quality control goals if a control rule intermediate between the 1_{2s} and 1_{3s} rules is employed. Specifically, using three control samples per batch, the $1_{2.385s}$ rule will yield a 5 percent false-rejection rate and a 90.6 percent rejection rate. In general, the best control rules in terms of control sample requirements are those for which the control limits have been calculated to achieve the specific quality control goals (Bishop and Nix 1993).

A somewhat different approach to evaluating control performance is taken when considering the detection of long-term, or persistent, quality degradation in a method. Here the focus is on how many batches will be accepted before the control rule leads to a batch rejection and, thereby, detection of the quality problem (Nix *et al.* 1987). The most informative way to present this performance behavior is as the cumulative run length distribution for the control rule. This distribution gives the probability of having rejected any batch, including the current batch, as a function of the number of batches run since the inception of the quality problem. Figure 2.6 shows the cumulative run length distributions for the 1_{2s} rule at five different levels of persistent method bias. The graph also has a line that indicates the medians of the distributions. For instance, the median run length for zero method bias is 15 batches. That is the median number of consecutive batches that can be expected to be accepted using this control rule when the quality of the method is unchanged. The median run lengths for the other levels of persistent method bias in the graph are 9, 4, 2, and 1, in order of increasing bias. (Note: average run length is the usual measure for the evaluation of control rules in the setting of a persistent problem in method quality. Because run length frequency distributions are highly right-skewed, average run lengths will be larger than median run lengths. For example, the average run length for the 1_{2s} control rule is 22 batches when the quality of the method is unchanged. Compare this to the median value of 15 batches. Because it is less influenced by extreme run length values of low probability, median run length better reflects the run length behavior expected of a control rule.)

The control rule used to monitor a method for persistent quality degradation must satisfy the clinical performance and false-rejection quality control goals for such problems. The performance goal can be expressed as maximum median (or average) run length at a specified level of quality decline in the method. The false-rejection quality goal can be expressed either as an acceptable median run length when the quality of the method is unchanged or, as for single-batch quality monitoring, as an acceptable false-rejection rate per batch. Typically, single rule control procedures are not able to provide the

Table 2.4
Multiple-Rule Control Procedure (SD is $SD_{within\text{-}laboratory}$)

Warning rule	
1_{2s}	one control result exceeds mean ± 2 SD
Within batch rules	
1_{3s}	one control result exceeds mean ± 3 SD
R_{4s}	one control result exceeds mean + 2 SD one control result exceeds mean – 2 SD
Within and between batch rules	
2_{2s}	two consecutive control results exceed mean + 2 SD or mean – 2 SD
4_{1s}	four consecutive control results exceed mean + 1 SD or mean – 1 SD
$10_{\bar{x}}$	ten consecutive control results fall on one or the other side of the mean

requisite combination of performance and false-rejection rate. A number of multiple-rule procedures have been proposed that perform much better than single-rule procedures. The multiple-rule procedure of Westgard *et al*. (1981) has achieved the greatest popularity. The rules used in the procedure are listed in Table 2.4. The control sample set for this procedure consists of a low concentration control sample and high concentration control sample. If both control results pass the 1_{2s} control rule using the respective established result distributions, the batch is considered to be in-control. If one of the control results exceeds its 2 $SD_{within\text{-}laboratory}$ limits, evaluate the results using the remaining control rules. If the results fail to pass any of the rules, the batch of measurement is rejected. If the results pass all of the rules, the batch is in-control.

Cumulative sum procedures and moving average procedures have also been proposed as tools for monitoring for persistent degradation in method quality (Nix *et al*. 1987, Strike 1996). The performance of these approaches in the control of method trueness has been shown to be superior to that the multiple-rule procedures (Bishop and Nix 1993, Parvin 1992). These approaches are computational intensive but can easily be implemented in the modern computerized laboratory.

Test sample-based quality control. The use of control material for internal quality control has several practical problems. The material is expensive, it may have limited stability, and often its composition is different from that of the test samples. This has prompted the development of alternative procedures for monitoring method precision and trueness that use test samples rather than control material. These procedures have not been accepted as replacements for quality control using control material but many laboratories use them to supplement their control material-based internal quality control program.

To monitor within-run precision, a test sample is divided into aliquots prior to being assayed and the aliquots are run in the same batch. The variance of the replicate results for the test sample is calculated using the formula,

$$var = \frac{\Sigma (x_i - mean)^2}{n - 1}$$

where x_i is the i th replicate result, *mean* is the mean of the replicates, and n is the number of replicates. If duplicates are used, the formula is,

$$var = \frac{(x_1 - x_2)}{2}$$

Once 20 to 30 such replicate determinations have been made, the within-run imprecision is estimated using the formula,

$$SD_{within-run} = \sqrt{\frac{\Sigma var_j}{N}}$$

where N is the number of test samples studied. This estimate is compared to the within-run imprecision established for the method. When the next test sample replicate set is run, its variance is added to the variance data set and the oldest variance value in the data set is deleted. The within-run imprecision is recalculated and the updated estimate is compared to the imprecision standard. Within-laboratory analytic precision can be evaluated in a similar fashion by having test sample replicates assayed in different batches.

To monitor method trueness, the mean value is calculated for a block of consecutive test sample results. The block may consist of a specified number of results, of all of the results in a batch, or of all of the results for a defined period of time, usually a day. The block mean is compared to the population mean established for the method. Assuming that the mix of patients is similar over time, the mean value of a block of test sample results will equal the established population mean. Truncation of the data to exclude extreme result values improves the performance of the method. Improved performance also comes from the use of a

smoothing function for the calculation of the mean values (Gardner 1985, Strike 1996). Smoothing leads to a reduction in the variability in the estimates of the means that would otherwise arise from day-to-day variation in the makeup of the clinical population from which the test samples come. In addition, if the test sample mean is recalculated for each successive result rather than for blocks of results, method trueness can be continuously monitored, although continuous monitoring may not offer any better performance than block-wise monitoring (Smith and Kroft 1997).

External quality control

External quality control refers to the procedures for quality control that involve the participation of two or more laboratories. Proficiency testing is an external quality control program mandated by a regulatory body for the purpose of determining laboratory quality. Most commonly, programs of external quality control are conducted by professional societies or manufacturers of control materials. Control material is provided to participating laboratories where control samples are assayed on a regular schedule, usually in conjunction with the internal control samples. All of the participating laboratories receive control material from the same production lot. The control sample results are transmitted to the program sponsors who analyze the data and issue reports that describe the result distributions among the participating laboratories and indicate the location of the individual laboratory results within that distribution. In this way, external quality control serves as an adjunct to internal quality control by providing an additional mechanism for monitoring the long-term trueness of a method.

There are a number of other important aims served by external quality control. These aims include to provide a measure of the "state of the art" for the measurement of an analyte; to obtain consensus values for control material when neither definitive nor reference methods exist for the measurement of an analyte; and, importantly, to investigate the sources of inter-laboratory variability in the measurement of an analyte (Büttner *et al*. 1983b). By analyzing subgroup results from a large number of laboratories, it is possible to compare the performance of different methods and to evaluate the extent to which inter-laboratory measurement variability is explained by laboratory variables such as laboratory size and laboratory workload.

METHOD PRACTICABILITY

Practicability refers to those properties of a method that relate to practical aspects of its implementation. These include speed, cost, technical skill requirements, dependability, and safety (Büttner *et al*. 1980a). These are clearly important concerns in the decision to employ a particular method.

The speed of a method is determined by the time needed for method and sample preparation, the time spent assaying the sample, and the time required for calculation of the results. The time needed for method preparation is at its longest if specimens are received infrequently and one-at-a-time, for the method must then be set up anew for each specimen. Method preparation is at its shortest if samples are received and run while the method is maintained in a fully operational state. There is sometimes a trade-off between the speed of a method and its quality. This may mean that two methods need to be set up in the same laboratory, a slower method of high quality that is used for routine work and a rapid method of lower quality that is used when circumstances demand a quick turnaround, provided, of course, that lower quality results can be tolerated clinically in exchange for rapidity in obtaining the results.

The cost of a method includes not only the expense of the reagents and materials used in sample preparation and assay, but also capital and operation costs, such as maintenance and repair costs for the instrument on which the method is implemented, labor costs, which will vary depending upon the technical expertise required of the staff who run the method, and overhead costs. Besides its affect upon method cost, the technical skill requirements of a method, which determines who on the staff can run the method, determines when it can be run—only when the appropriate staff members are scheduled to be at work.

The dependability of a method quantifies the rate or frequency with which the method succeeds in producing valid results. Dependability is not wholly intrinsic to the method but can vary from location to location due to differences in laboratory conditions and staff quality. This is particularly relevant when the method is to be used outside of the central clinical laboratory, in a setting closer to the patient. Such point-of-care testing may take place in a satellite laboratory, on a hospital ward, in a physician's office, or in the home of the patient. In these

settings, the personnel are typically trained in the performance of the method but usually have only a limited amount of training and experience in general laboratory practices. Consequently, the method needs to be easy to perform and highly reliable if it is to be dependable. For home testing, in which the patient or a family member performs the test, the need for method ease and reliability is even greater.

Method safety encompasses the whole range of safety considerations in the performance of a method. It includes concerns for the biological, chemical, and radiation hazards to which the laboratory staff may be exposed during the preparation of samples and the performance of the method and for the electrical and mechanical safety of equipment used in the performance of the method. Safety considerations may determine who can run the method and where the method can be set up in the laboratory.

METHOD EVALUATION

There are two kinds of method evaluations: the evaluation of a newly developed method, which is usually undertaken and reported by the laboratory scientists who devised the method, and the evaluation of a validated method, which is performed by the laboratorians who are considering setting up the method in their clinical laboratory. The first type of evaluation is ordinarily conducted under the best possible laboratory conditions and with the loving attention of the developers. The second type of evaluation, which is usually conducted under routine laboratory conditions, establishes how well a method performs in the laboratory in which it will be used. The performance may not be as good as that reported by the developers because of the differences in the operating conditions. That is why on-site method evaluation is necessary before implementing any method. The evaluation of a newly developed method must be very thorough. The report of the evaluation should include the components listed in Table 2.5.

Laboratory methods are developed for three reasons. The first is to provide a method for the measurement of an analyte which is recognized to be clinically useful but for which there is no existing method. The second is to provide a method with analytic quality superior to that of other methods currently in use and the third is to provide a method of greater practicability than that of currently available methods. In an exemplary evaluation of a new method for determining plasma phosphate concentration, Luque de Castro *et al.* (1995) indicate that their primary motivation for developing the method was to improve practicability. The method uses a flow injection (FI) system with immobilized enzymes. As they state, immobilized enzymes

> have several advantages over the use of dissolved enzymes in batch assays, such as lower analytical cost, higher selectivity and stability, and long life span.

A similar method had already been developed,

> Male and Luong (16) developed the first FI method for the determination of phosphate with immobilized [nucleoside phosphorylase] and xanthine oxidase

but that method used amperometric detection of the reaction product. The method of Luque de Castro *et al.* produces a different indicator product that is detected fluorometrically.

Table 2.5
Components of a Method Evaluation Report

1. Statement of the motivation for the development of the method
2. Description of the method
3. Description of the optimization of analytical variables
4. Characterization of the calibration curve
5. Assessment of analytical quality
6. Determination of analytical range

Method description

The description of the method should include the same items as are found in a written measurement procedure (Table 2.3). The level of detail should also be comparable to that of a written procedure: using only the method description and general laboratory knowledge, a reader should be able to setup and perform the method in his or her clinical laboratory.

Sampling and specimen handling. The report should state the kinds of specimens that can be assayed using the method. It should be noted if special processing is required of some specimens. For example, urine may need to be diluted prior to measurement. The kinds of specimens actually used in the performance of the method evaluation should be indicated.

Principle and method of measurement. The statement of the principle of measurement of the method should include the techniques of analyte separation and of signal generation and detection. The phosphate method of Luque de Castro *et al.* does not employ an analyte separation step. The signal is generated by an enzymatic reaction for phosphate coupled to the production of an fluorescent end-product,

> A number of enzymatic methods . . . are based on the ability of phosphate to activate some enzymatic reactions catalyzed by glyceraldehyde-3-phosphate dehydrogenase (8), phosphorylase *a* (9), maltose phosphorylase (10), sucrose phosphorylase (11), and nucleoside phosphorylase

The method of Luque de Castro *et al.* utilizes nucleoside phosphorylase as the enzyme,

$$\text{Inosine} + P_i \overset{\text{NP}}{\rightleftharpoons} \text{hypoxanthine} + \text{ribose-5-P}$$

$$\text{Hypoxanthine} + 2\,H_2O \overset{\text{XOD}}{\rightleftharpoons} \text{uric acid} + 2\,H_2O_2$$

$$\text{p-Hydroxyphenylacetic acid} + 2\,H_2O_2 \overset{\text{HPOX}}{\rightleftharpoons} H_2O + \text{bi(p-hydroxyphenylacetic acid)}$$

where NP is nucleoside phosphorylase, XOD is xanthine oxidase, and HPOX is peroxidase.

> The species monitored fluorometrically is the dimer of *p*-hydroxyphenylacetic acid (*p*-HPA), which exhibits maximum excitation at 325 nm and maximum emission at 415 nm.

In this method analyte specificity is achieved both by the use of an enzymatic reaction specific for phosphate and by the use of fluorometric detection.

Reagents and apparatus. As in the written procedure, the method evaluation report should stipulate the identity and source of reagents and describe the preparation of stock solutions and working solutions. The storage conditions and shelf lives of the solutions should be stated. The preparation of the immobilized enzyme reactors (IMERs) is of particular note in the method of Luque de Castro *et al.*,

> NP and XOD were immobilized on controlled-pore glass (CPG 120-200 mesh; Electronucleonics, Fairfield, MA) by using Masoom and Townshend's procedure (17). Pump tubes of different lengths [1.5 mm (i.d.)] were then packed with each support-enzyme conjugate and stored at 4°C in the following solutions: 100 mmol/L Tris-HCl, pH 7.0, for the NP immobilized enzyme reactor (IMER) and 1 mol/L ammonium sulfate + 0.5 mmol/L sodium salicylate in 100 mmol/L Tris-HCl, pH 7.0, for the XOD IMER. Under these conditions both enzyme reactors kept their activity for at least 3 weeks.

The description of the instruments used in the method should include a detailed discussion of any modifications required for the performance of the method. In the case of flow injection systems, such as that of Luque de Castro *et al.*, the flow injection manifold needs to be described.

> The hydrodynamic system used (Fig. 1) consists of a peristaltic pump that propels the reagent streams through the channels. The sample, diluted appropriately, is injected into a stream of reagent A, which merges with a stream of reagent B; reagent B contains inosine, the substrate for NP biocatalysis. The first two enzymatic reactions take place along the NP and XOD IMERs. An additional merging point located after the IMERs allows the main stream to be mixed with reagent C (which contains *p*-HPA and HPOX), which reacts and catalyzes, respectively, the derivatizing reaction of the hydrogen peroxide produced in the previous step. The derivatizing reaction is developed along the reactor. Finally, the sample reaches the flow cell and provides the analytical signal. The enzyme reactor and the open reactor are thermostated at 37°C.

Preparation and use of the measurement system. In most cases, the preparation and use of the measurement system is a matter of general laboratory knowledge so no explicit discussion of these topics is required. If the instrumentation is new or if a familiar instrument is modified or operated in a novel fashion, the report should

provide a detailed, preferably stepwise, account of the performance of the method.

Analysis and optimization of analytical variables

The principal objective in the development of a laboratory method is to end up with the maximum possible analytical quality given the level of practicability envisioned by the developers. To achieve this goal, it is necessary to identify the optimal combination of operating set-points for the analytical variables of the method. This requires an analysis of the dependency of the quality endpoints upon the operating set-points. Luque de Castro *et al.* performed an exceptionally thorough analysis of this sort in the development of their method. They used the magnitude of the analytical signal as the primary quality endpoint and studied a variety of analytical variables,

> The variables affecting the analytical process and hence the signal it provided were classified as chemical, physical, and hydrodynamic (Table 1), and then studied by univariate analysis.

Table 1. Assay variables.

Variables	Range studied	Optimal value
Physical		
Temperature, °C	20–45	37
Chemical		
Tris-HCl buffer, mmol/L	50–500	100
pH	6–9	8.5
Inosine, mmol/L	1.0–7.5	4.75
p-HPA, mmol/L	2–15	10
HPOX, U/L	4–12	8
FIA		
Flow rate, mL/min	0.3–2.5	1.60
Injected volume, μL	50–500	300
Length of L_1, cm	50–400	250
Length of NP-IMER, cm	0.5–3	1
Length of XOD-IMER, cm	0.5–3	1

The analysis and optimization of the flow injection variables is described as follows,

> High flow rates (2.32 mL/min) decreased the analytical signal, but low flow rates (0.58 mL/min) decreased the sampling frequency and resulted in increased dispersion. A flow rate of 1.60 mL/min was selected as a compromise.
>
> A sample volume of 300 μL was chosen to obtain the best analytical signal, since at greater volumes the signal remained almost constant.
>
> The optimal lengths of the enzyme reactors were 1 cm each. Using a longer NP IMER provided a sharp increase in the baseline and a decreased analytical signal. Increasing the XOD IMER did not improve the analytical signal.
>
> A length of 250 cm for the open reactor was enough to achieve a reproducible mixture of reagent C and the main stream, thus providing optimal analytical signal.

Notice that, even though the primary quality endpoint was the magnitude of the analytical signal, the flow rate that was selected as optimal was not the flow rate resulting in the maximum value of the signal. Larger signals were obtained at lower flow rates. However, the lower flow rates decreased measurement precision, a secondary quality endpoint, and decreased the sampling frequency of the system, a practicability endpoint. It is not infrequent in method development that the choice of an optimal analytical variable set-point represents such a compromise among competing quality and practicability considerations.

A univariate approach to method optimization was employed by Luque de Castro *et al.* In the univariate approach, the response of a system to the set-point of one variable is studied with all of the other variable set-points held constant. This works fairly well if all of the variable set-points are held near to their true optimal values or if the sensitivity of the system to the set-point of each variable is largely independent of the set-points of the other variables. It works poorly if the approximate values of the optimal set-points are not known beforehand and if there is interdependence among the variables in their effect upon the system response.

A multivariate optimization approach can be used in circumstances in which the univariate approach is not likely to perform well. In the multivariate approach, none of the set-points of the analytical variables are kept constant; instead, the response of a system to various set-point combinations is studied (Box and Draper 1987). The combinations are chosen so that they will cover what is *a priori* believed to be the most interesting portion of the multivariate solution space. This provides data points on the response surface, the multi-dimensional surface that characterizes the relationship between system response and the set-points of the analytical

variables. These points can be used to fit a response surface model, if a functional form for the surface is suggested by the data. The optimal set-point combination can then be calculated from the model equation. Alternatively, the data points can be used as a starting point for an empirical search algorithm. Search algorithms seek out optima in an iterative fashion: using the response data from the preceding step, the algorithms indicate the most informative set-point combinations to test next. The iterations continue until the maximum system response and its associated set-point combination is identified. The popular search algorithms are highly efficient and rapidly converge to the response surface optimum. This means that the delineation of the optimal analytical variable set-point combination for a new method can be achieved without undue expense. Multivariate optimization should, therefore, be considered whenever there is uncertainty about the validity of the univariate optimization approach.

Characterization of the calibration curve

The most common form for the equation of the calibration curve is a straight line. Sentiments of the sort, "Linearity is a state sought by all clinical laboratorians. It means straight and predictable, good work, and good value" (Passey and Maluf 1992) are common despite the fact that there are measurement systems in the clinical laboratory, such as competitive immunoassay systems, that produce results of high quality despite having nonlinear calibration curves. Nevertheless, when calibration linearity is possible, considerable efforts are made to assure that the operating conditions for a measurement system yield a linear calibration curve.

In order to evaluate the linearity of a calibration curve, a measure or test of linearity is needed. Such a measure has not proven easy to come by. Tholen (1992) listed 22 different statistical techniques that had been proposed for evaluating calibration linearity up to the time he wrote his review. Since then, a technique developed by Kroll and Emancipator (1993, Emancipator and Kroll 1993) has achieved a degree of acceptance in the laboratory medicine community as a linearity measure. That technique is based upon the quantification of nonlinearity as the root mean square of the deviation of the calibration curve from an ideal straight line,

$$nonlinearity = \sqrt{\frac{\int_{x_{low}}^{x_{high}} (c(x) - g(x))^2 dx}{x_{high} - x_{low}}}$$

where $c(x)$ is the equation of the curve that best fits the empirical calibration data, $g(x)$ is the equation of the ideal straight line fit of the data, and x_{high} and x_{low} are the values of the high and low calibrators, respectively. Defining the relative nonlinearity as,

$$relative\ nonlinearity = \frac{nonlinearity}{y_{high} - y_{low}}$$

where y_{high} and y_{low} are the highest and lowest signal magnitudes recorded during the linearity study, Emancipator and Kroll (1993) found that calibration curves that are acceptably linear by visual inspection have relative nonlinearities of less than 2.5%. Using this technique, Luque de Castro *et al*. found that their phosphate method demonstrated acceptable linearity over the range of measurement,

> The linearity of the method was assessed by means of Kroll and Emancipator's procedure (18,19) recently adopted by the College of American Pathologists (20). the . . . nonlinearity was 0.063 mmol/L; the relative nonlinearity, 1.29%.

A number of practical considerations are involved in the implementation of the technique of Kroll and Emancipator. The number and spacing of the calibrators need to be chosen. Kroll and Emancipator suggest using 5 equally spaced calibrators. This scheme will reveal both monotonic and sigmoidal nonlinearity. In cases in which the curvature in the calibration curve appears to be limited to one or the either end of the curve, such as when there is concavity at the high end of a calibration curve for a method which suffers substrate exhaustion at high analyte concentrations, it may be necessary to add additional calibrators within the suspect interval. Each calibrator should be run in replicate. The number of replicates needed depends upon the measurement variability of the method: for highly precise methods, duplicates are adequate; for methods that are imprecise, quintuplicates are warranted. Most authors report the use of triplicates, a convenient compromise number. The replicates should be between-run rather than within-run (Kroll and Emancipator 1993, Emancipator and Kroll 1993).

The formula for nonlinearity requires two equations, one for the curve that best fits the empirical calibration data and one for the ideal straight line fit of the data. Kroll and Emancipator recommend, and themselves use, polynomial equations to derive the best fit curve. Polynomial equations are a good choice because they can be readily fit to the data by weighted multiple linear regression and they are also easy to integrate. The ideal straight line fit is the line that results in the minimum value for nonlinearity. Formulas for calculating the slope and intercept of this line can be found in Emancipator and Kroll (1993). Note that these equations refer to the calibration curve and, as such, relate the calibrator concentration to the signal magnitude. The y data in a linearity study, therefore, are signal magnitudes. Luque de Castro *et al.* properly used the strength of the fluorescence signal as the y variable in the linearity study of their method,

> A series of eight standard solutions with concentrations between 0.1 and 20.0 μmol/L were prepared from the phosphate solution described in *Materials and Methods*. The equation of the analytical signal obtained by triplicate injection of these standards into the FI manifold was as follows: fluorescence intensity = 27.5 + 49.2 [P_i] (μmol/L)

Some authors report using measurement results rather than signal magnitudes as the y data. This is improper and, even more to the point, paradoxical in that the calibration curve is the matter under study; without a calibration curve there cannot be a measurement curve and, in turn, there cannot be measurement results.

Assessment of analytical quality

The quality of an analytical method is assessed through a characterization of the method trueness and the method precision.

Trueness. As discussed earlier, trueness is the closeness of agreement between the true value of an analyte and the average value obtained from a large number of replicate measurements. To evaluate the trueness of a method it is, therefore, necessary to know the true analyte value in a sample. Indeed, it is necessary to know the true value in a number of samples with analyte concentrations that vary across the proposed range of measurement of the method. This requirement can be met in either of two ways. If the method evaluators have access to a reference method for the measurement of the analyte, the true concentrations can be determined in clinical samples from the evaluators' laboratory. Otherwise, the evaluators can use certified reference material for which the true concentration of the analyte of interest have been determined by the certifying agency through the use of reference or definitive methods.

The steps in the characterization of the trueness of a laboratory method are depicted in Figure 2.7. Between-run replicate measurements are made of the samples with known analyte concentration (top graph). At least 5 replicate measurements should be performed and 10 is preferred. The average value of each set of replicates is computed and the differences between the averages and the true values are calculated. The differences, which represent the bias of the method at the sampled analyte concentrations, are plotted (middle graph). In order to characterize the bias at all concentrations within the range of measurement, a linear bias model is fit to the data using weighted linear regression. Bias is classified as being constant if the constant term of the model is nonzero. It is classified as proportional if the slope of the model is nonzero. In the example, the bias shows a mixed pattern.

The graph of the bias model is referred to as a bias profile (Keller and Passing 1989). The bias profile can be graphed in terms of absolute bias (middle graph) or relative bias, which is the ratio of the bias to the analyte concentration expressed as percent (bottom graph). It is especially useful to graph the bias profile in relative terms because method bias criteria, which are rules for deciding if a method shows adequate trueness for clinical purposes, are usually expressed in relative rather than absolute terms. A bias criterion can therefore be plotted on the same graph as the bias profile and the interval over which the method meets the criterion can be easily appreciated. An example of this is shown in the bottom graph. The bias criterion depicted is a relative bias of less than 10%. The relative bias of the method satisfies the criterion at analyte concentrations greater than 19 units.

Recovery. If method evaluators do not have access to a reference or definitive method and if there are no readily available certified reference materials, method trueness cannot be evaluated by the approach outlined in the preceding section. Trueness must then be evaluated by means of a recovery study. A known amount of analyte is

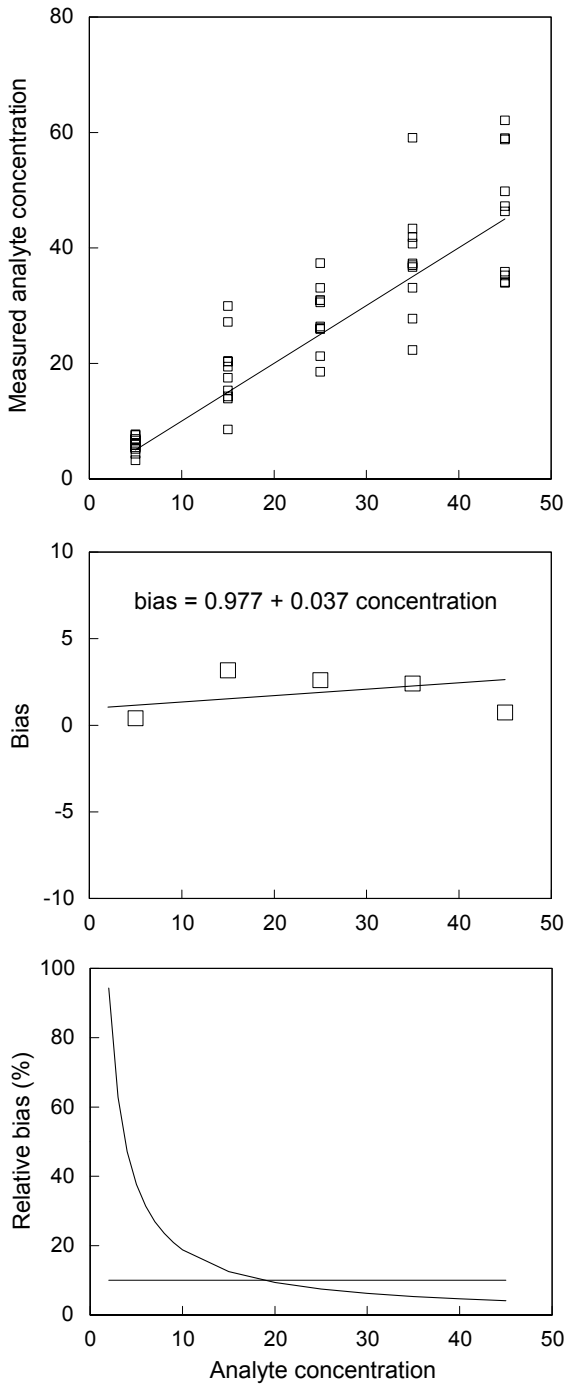


Figure 2.7 The bias of a hypothetical laboratory method. The top graph shows replicate measurement data (symbols) and the line of identity. The middle graph shows the empirical biases (symbols) and the linear model fit (line). The bottom graph shows the relative bias profile (curve) and a bias criterion (line).

added to a clinical sample and the increment in analyte concentration as measured by the method is compared to the increment in concentration calculated from the sample volume and amount of analyte added. The comparison is usually expressed in terms of the percentage recovery,

$$\text{recovery} = \frac{\text{measured concentration increment}}{\text{predicted concentration increment}} \times 100\%$$

Using multiple aliquots of a clinical sample with a low initial analyte concentration, a number of recovery samples of varying final concentrations are made. The concentrations should span the proposed range of measurement of the method. Between 5 and 10 between-run replicate determinations are performed on each of the recovery samples and the average recovery at each addition level is calculated. Luque de Castro *et al*. used a recovery study to evaluate trueness in their method evaluation,

> Two aliquots of six samples were subjected to additions of standards (0.5 and 1.7 mmol/L) to establish the recovery of the method. The results obtained (Table 3) ranged from 96% to 104%, which represent a good recovery for the supplemented samples.

Table 3. Assay recovery.

		Addition 1[a]		Addition 2[a]	
Sample	Phosphate, mmol/L	Found, mmol/L	Recovery, %	Found, mmol/L	Recovery, %
1	1.28	1.78	100	2.94	98
2	1.79	2.30	102	3.50	100
3	1.56	2.04	96	3.26	100
4	0.45	0.92	98	2.18	103
5	1.20	1.69	98	2.93	101
6	1.02	1.54	104	2.73	100

[a] 0.5 and 1.7 mmol/L for additions 1 and 2, respectively.

There are two potential problems with recovery studies that should be kept in mind. First, the calculated increment in concentration in a recovery sample is subject to error due to possible errors in the amount of analyte added or in the measurement of the volume of the sample. Second, unless a clinical sample with a very low analyte concentration can be found, the trueness of the method is studied only in the higher concentration range (initial analyte concentration plus added analyte concentration).

Cross-reactions and interferences. Method trueness is also evaluated by studying the effects of potential interfering and cross-reacting substances. Usually the substances are studied one at a time. The potential interferent or cross-reactant is added to a clinical sample and the change in the analyte concentration is measured. The effects of the

substance are most often expressed in terms of percent change in measured analyte concentration. The substance is evaluated over the range of substance concentrations that spans the anticipated clinical range for the substance so that the effects can be related to the concentration of the substance, usually by use of a linear model. This one-at-a-time approach that was taken by Luque de Castro *et al.*,

> The effects of bilirubin and hemoglobin as potential optical interferences and of uric acid, ascorbic acid, xanthine, hypoxanthine, and glucose as possible chemical interferences were studied. Each compound was added to a pool of serum (phosphate concentration 1.25 mmol/L) and its influence established. No interferences were detected from bilirubin for concentrations $<$600 mg/L, hemoglobin $<$20 g/L, uric acid $<$150 mg/L, ascorbic acid $<$75 mg/L, and glucose $<$200 mmol/L . . . The presence of allopurinol $<$ 200 mg/L in serum did not cause interference; it did cause an error of $\sim$ -4% at 400 mg/L.

Xanthine and hypoxanthine, which are substrates of the second enzymatic reaction, produced a positive interference, which presumably was proportional to the concentration of the added substance. The authors comment that interference from these substances in the clinical setting should be "no special problem" because they are "present in serum very infrequently."

The one-at-a-time approach will not reveal complex chemical interactions such as those occur between an interferent or cross-reactant and the analyte and those that occur between different interferents and cross-reactants. Quantitative evaluation of complex interferences and cross-reactions requires a response surface modeling approach similar to that discussed in the section on optimization of analytical variables (Kroll and Chesler 1992). In this approach, the potential interferents and cross-reactants are added in varying concentrations to each aliquot of the clinical sample. For a complete factorial design, all possible combinations of the various concentrations for each of the interferents and cross-reactants are studied (Box and Draper 1987). The measured analyte concentrations are fit to an response surface model by multiple regression analysis. A model limited to first-order and interaction terms is usually used, such as the following ,

$$\textit{analyte concentration} = b0 + b1x1 + b2x2 + b12x1x2$$

for two interferents where *x1* and *x2* are the concentrations of the respective interferents and *b0* represents the concentration of analyte in a sample free of interferents.

The presence of complex chemical interactions is indicated by statistically significance of the coefficients of the interaction terms. The clinical significance of an interaction depends upon the magnitude of the interaction effects.

Precision. Figure 2.8 illustrates the steps in the characterization of the precision of a method. Replicate measurements are made using samples with analyte concentrations that span the measurement (top graph). If within-run imprecision is being studied, all of the measurements on a sample must be made during the same run. In a study of within-laboratory imprecision, the measurements on a sample need to be performed during different runs and, preferably, on different days. An absolute minimum of 10 replicate measurements need to be made to obtain a moderately precise estimate of the standard deviation; 20 replicates is better and 50 replicates is better still (Sadler and Smith 1990).

The standard deviation of each set of replicate measurements is calculated using the formula (Bookbinder and Panosian 1986),

$$\textit{standard deviation} = \sqrt{\frac{\Sigma (x_i - \textit{mean})^2}{n-1}}$$

where x_i is the i th replicate result, *mean* is the mean of the replicates, and n is the number of replicates. The standard deviations are plotted versus analyte concentration. If the imprecision is constant over the measurement range, the empirical standard deviations will be roughly equal and will form a fairly flat line. If the imprecision is proportional to analyte concentration, the empirical standard deviations will increase in magnitude with increasing analyte concentration. The imprecision model that is usually fit to the empirical data is the 3-parameter model proposed by Sadler *et al.* (1988),

$$SD = (b0 + b1\ \textit{concentration})^{b2}$$

This model is quite flexible. If *b2* is one, the model defines a line. Otherwise, the model defines a curve that can be either convex (*b2* greater than one) or concave (*b2* less than one). Because the model is nonlinear, it must be fit by nonlinear regression.

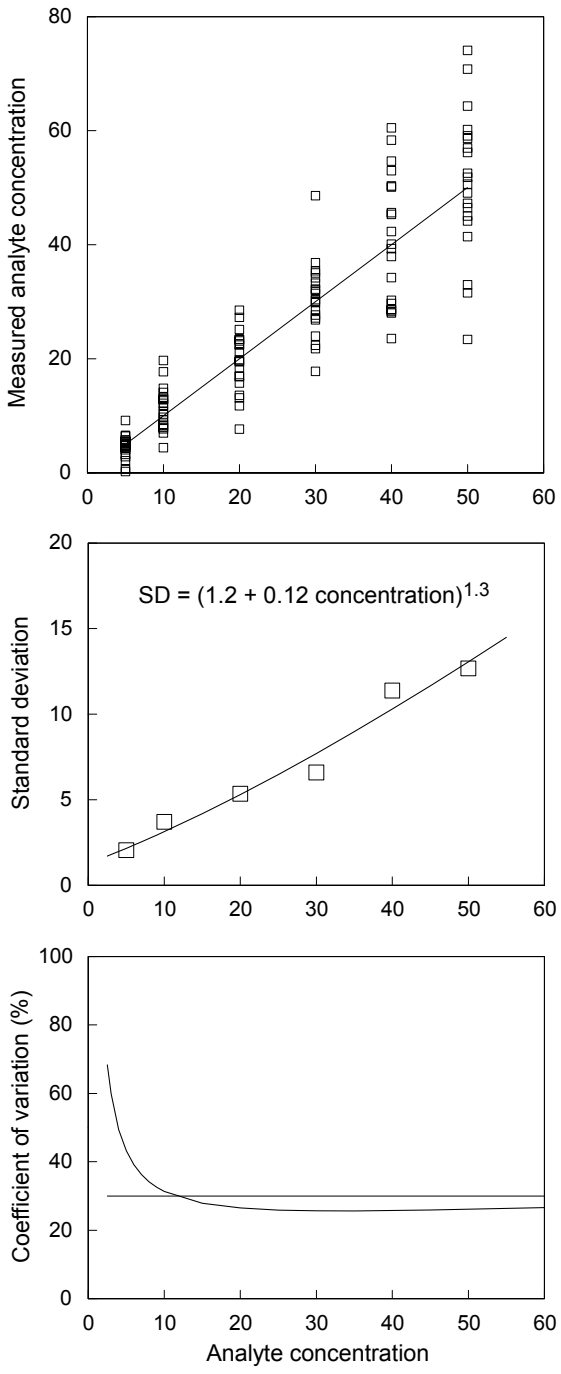


Figure 2.8 The precision of a hypothetical laboratory method. The top graph shows replicate measurement data (symbols) and the line of identity. The middle graph shows the empirical standard deviations (symbols) and the imprecision model fit (curve). The bottom graph shows the relative imprecision profile (curve) and a precision criterion (line).

The graph of the imprecision model was originally called a precision profile (Ekins 1983) but the designation, imprecision profile, has become more popular. The y-axis of the imprecision profile can be either the magnitude of the standard deviation (middle graph) or the magnitude of the coefficient of variation (bottom graph). Almost always, imprecision profiles are graphed with imprecision quantified as coefficient of variation. The reason is similar to that for graphing the bias profile in relative terms: method precision criteria are expressed in relative terms, i.e. as coefficients of variation, so they can be plotted on the same graph as the imprecision profiles. A method precision criterion of 30% is plotted in the bottom graph. The precision of the method satisfies the criterion at analyte concentrations greater than 11 units.

In the phosphate method evaluated by Luque de Castro *et al.,* within-run and between-run precision were evaluated at three different concentrations,

> The precision . . . of the method was checked by assaying three serum pool samples from a clinical laboratory that contained low, medium, and high concentrations of phosphate (~0.900, 1.070, and 1.700 mmol/L, respectively). Aliquots of the three samples were analyzed after a 1:250 dilution, both in single run and during 11 days for within- and between-run studies, respectively.

Table 2. Assay precision.

	Phosphate, mmol/L		
	Mean	SD	CV, %
Within-run (n = 22)			
Level I	0.890	0.018	2.03
Level II	1.064	0.016	1.50
Level III	1.700	0.012	0.70
Between-run (n = 22)			
Level I	0.909	0.031	3.3
Level II	1.073	0.021	2.0
Level III	1.669	0.017	1.7

Because the range of measurement is small, three concentrations seems an adequate number to study. Typically, the range of measurement is much larger and a greater number of concentrations need to be studied. The authors did not model the precision data even though neither within-run nor between-run precision appear to be constant.

The preceding discussion and example describe the separate evaluation of within-run and between-run precision. This is a straightforward and clear-cut way to conduct a precision study but it isn't the only way that such studies are performed. Often, a few within-run replicates are assayed in each of a large number of runs and both within-run precision and between-run precision are computed from the

resulting data set. The variance of the replicate results for each run is calculated using the formula,

$$var_j = \frac{\Sigma (x_i - mean_j)^2}{n-1}$$

where x_i is the i th replicate result in run j, $mean_j$ is the mean of the replicates in run j, and n is the number of replicates assayed per run. If only two replicates are assayed per run, n-1 is set equal to 2 not to 1. This biases the analysis somewhat, so it is better to assay three or more replicates per run. Within-run variance is calculated as the average of the individual run variances so,

$$SD_{within\text{-}run} = \sqrt{\frac{\Sigma var_j}{N}}$$

where N is the number of runs. The variance of the individual run replicate means is computed using the formula,

$$var_{means} = \frac{\Sigma (mean_j - overall\ mean)^2}{N-1}$$

where *overall mean* is the mean value of the individual run replicate means. Then (Box *et al.* 1978),

$$SD_{between\text{-}run} = \sqrt{var_{means} - \frac{SD^2_{within-run}}{n}}$$

A point that needs to be mentioned in any discussion of method precision is that within-run precision can always be improved by measuring samples in duplicate or triplicate and reporting the average value of the results. With replicate measurements, the imprecision decreases by a factor equal to the square root of the number of replicates. For duplicate measurements, the within-run imprecision is 0.71 times as large as with single measurements and with triplicate measurements, it is 0.58 times as large. This approach is costly in that fewer samples can be run per batch but the resultant improvement in method precision may significantly increase the clinical utility of the method.

Resolving power and detection limit. The resolving power of a method is expressed in terms of the minimum distinguishable difference in concentration, D_{min}. Using the formula,

$$D_{min} = z_c \sqrt{2}\, SD_{within\text{-}laboratory}$$

the resolution profile for a method can be calculated from the imprecision model, giving,

$$D_{min} = z_c \sqrt{2}\, (b0 + b1\ concentration)^{b2}$$

where the model parameters *b0*, *b1*, and *b2* apply to within-laboratory imprecision.

The detection limit can be calculated as the smallest concentration that solves the equation

$$concentration = z_c \sqrt{2}\, (b0 + b1\ concentration)^{b2}$$

An approximate solution can be found graphically from a plot of the resolution profile. It is the concentration at which the resolution profile intersects the line of identity. Using this starting value, the exact solution can be found using an iterative root solving algorithm such as Newton's method (Sadler *et al.* 1992). Such algorithms are now widely available in computer spreadsheet programs.

An alternative and more direct way of calculating the detection limit is to use the imprecision data from the replicate sets with concentrations that are near the detection limit. Typically, the zero concentration replicate set and the lowest concentration replicate set are used. The variances of the data sets are calculated (var_{zero} and var_{low}) as is the pooled variance,

$$var_{pool} = \sqrt{\frac{n_1 var_{zero} + n_2 var_{low}}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the number of replicates in the zero concentration and low concentration data sets, respectively. The detection limit is calculated using the following formula (Rodbard 1978, Büttner *et al.* 1980b),

$$detection\ limit = blank - t_c\, var_{pool} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where *blank* is the mean value of the zero concentration replicate set and t_c is the confidence coefficient as found with the t distribution. For a 95% confidence level and equal numbers of replicates in each data set, t_c equals 1.734 for 20 replicates, 1.701 for 30 replicates, 1.686 for 40 replicates, and 1.677 for 50 replicates.

Analytical range

The analytical range, or working range, is the range of analyte concentrations for which the method satisfies all of the following criteria: the bias of the method is within acceptable limits, the precision of the method is within acceptable limits, and, if appropriate, the calibration curve is acceptably linear. The range is determined by reference to the bias profile, the imprecision profile, and the findings of the linearity study.

It is obviously desirable that the analytical range cover the entirety of the pathophysiological range of values for the analyte measured by the method. If the analytical range falls short, explicit rules need to be devised to deal with samples that have analyte concentrations outside of the analytical range. In practice, this problem is almost always due to samples with analyte concentrations that are above the upper limit of the analytical range. The question in such cases is if appropriate dilution of the sample will yield a valid measurement. If so, the dilution procedure and the means for calculating a corrected result need to be included in the method procedure. If not, the manner of reporting the out-of-bound result needs to be specified in the procedure.

METHOD COMPARISON

The purpose of a method comparison is to ascertain if the test results for a set of clinical specimens as obtained from one field method are, on average, the same as those obtained by another field method. The comparison amounts to a exploration of the clinical equivalence of the two methods. Clinically equivalent methods can be freely substituted for one another. The substitution of a method for another with which it is not clinically equivalent requires the establishment of a new reference interval for the analyte being measured and the development of a conversion formula that can be used to equate test results from the old method with test results from the new study.

Motivation

A complete report of a method comparison includes the components listed in Table 2.6. The first component is a statement of the motivation for the comparison. The most common motivation is the contemplated replacement of a method with one that possesses greater practicability.

The following excerpt from Turpeinen *et al.* (1995) explains their motivation for undertaking a comparison of three different methods for measuring hemoglobin A_{1c} (HbA_{1c}). HbA_{1c} is a stable adduct of glucose and hemoglobin A. The percent of hemoglobin present as HbA_{1c} depends upon the blood glucose concentrations to which the hemoglobin is exposed over the life-span of the red cell and is, therefore, a useful clinical marker of long-term blood glucose concentration control in patients with diabetes.

Table 2.6
Components of a Method Comparison Report

1. Statement of the motivation for the comparison of the methods
2. Description of the analytical methods
3. Description of the study population
4. Evaluation of concordance of the methods
5. Assessment of clinical equivalence

> . . . the methods currently used for [HbA_{1c}] measurement in clinical chemistry laboratories show large differences between reported values, and comparison of results from different laboratories is difficult.
>
> At present there is no accepted standard or acknowledged reference method. Recently, calibration based on a cation-exchange HPLC method has been shown to increase the comparability between various analytical methods (5,6).
>
> In this study we compare our own high-resolution HPLC cation-exchange method (PolyCAT A) with two other assays: a boronate affinity binding assay (IMx) and an automated system for [glycohemoglobin] analysis by cation-exchange chromatography (Diamat ™).

The motivation for this method comparison is to compare two commercially available methods for the determination of HbA_{1c}. In addition, the methods are compared with a high-resolution version of the methodology that has been gaining support as a calibrating method, cation-exchange chromatography. In this instance, it can be appreciated that the employment of cation-exchange as a calibrating method can be taken as tacit acknowledgment that it is of adequate accuracy for routine laboratory practice and, indeed, is probably more accurate than most other methods in routine use.

Analytical methods

When comparing methods, it is of course essential that it be clear exactly what the methods are. It therefore behooves the laboratorian to thoroughly describe the methods under study. In fact, it is appropriate to provide the same degree of completeness in the method description as one would for a method evaluation. However, it may be possible to

cite a previously published description of the method or to refer to the manufacturer's instructions when a commercial method is being studied. When any options exist in the performance of a method, the options chosen should be indicated.

Study population

Method comparison studies are performed using clinical specimens that have been submitted to the laboratory in the course of the medical care of patients. Most often, the specimens that are used are those that have been submitted for determination of the analyte measured by the methods under study. This is an obvious necessity in the case of xenobiotics but it also makes sense for other kinds of analytes because the range of values for the analyte is usually largest among those patients in whom the analyte is being measured. For instance, HbA_{1c} concentrations are only measured in patients with diabetes. These patients have values that range from normal to greatly increased with the majority being slightly to moderately elevated. A random sampling of laboratory specimens would be expected to uncover only a few specimens with elevated concentrations. In keeping with this logic, in their comparison study, Turpeinen *et al.* utilized specimens obtained from patients with diabetes:

> For method comparison we used 123 blood samples obtained mainly from diabetics sent to our routine laboratory for HbA_{1c} analysis.

A wide clinical spectrum should be represented by the specimens used in a method comparison study. This is important for two reasons. First, the range of values of the analyte often relates to the spectrum of disease in the patients from whom the specimens are obtained. For instance, if the specimens that are submitted to the routine laboratory for HbA_{1c} analysis come almost exclusively from outpatients whose disease is well controlled, the values will be for the most part normal or slightly increased. The concordance of the methods in this range may not reflect the concordance at the high values seen in patients whose disease is out-of-control. The HbA_{1c} values reported by Turpeinen *et al.* in their article include many above 10% of total hemoglobin, indicating moderate to severe chronic hyperglycemia, so their clinical population clearly encompasses a broad spectrum of glycemic control. The second reason to seek a broad clinical spectrum is to guarantee a wide spectrum of biochemical variability. The wider the biochemical spectrum, the more likely it is that measurement differences due to differential method specificity will be detected.

Evaluation of concordance

There are two general approaches for the evaluation of concordance between the results of field methods, regression analysis and difference analysis. Correlation analysis is not a useful approach for evaluating concordance for a number of reasons (Bland and Altman 1986, Hollis 1996). First and foremost of these is that the correlation coefficient is not a measure of agreement between data pairs but, rather, is a measure of the goodness-of-fit of a linear model of the data pairs. Thus, for example, data pairs that are aligned along the line,

$$result_{method\ 2} = 10 + 2\ result_{method\ 1}$$

will have a perfect correlation coefficient despite the fact that the data pairs do not agree at all. Another problem with the correlation coefficient is that its value depends upon the range of the data analyzed. The wider the range, the larger the correlation coefficient. In this way, the inclusion of extreme data pairs, even pairs with less than average agreement, will inflate the value. Yet another problem with the correlation coefficient is that it reflects data variability as well as data linearity. As a result, the correlation coefficient of two highly precise methods that, on average, do not agree particularly well may be larger than the correlation coefficient of two less precise methods that, on average, agree very well.

Regression analysis. The goal of regression analysis, which has been the standard approach for concordance evaluation for decades, is to define the functional relationship between the results of the methods so that it can be compared to the relationship that characterizes ideal concordance. In practice, this means using a linear regression technique to find the equation of the line that best fits the paired result data,

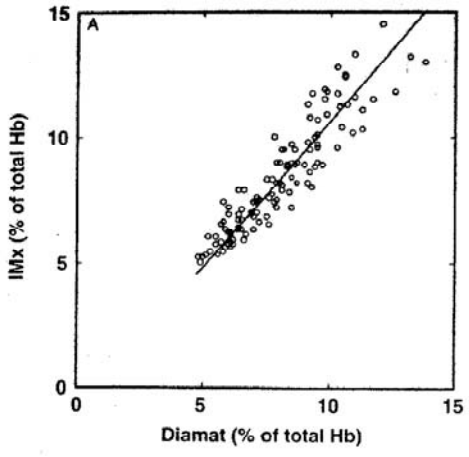
$$result_{method\ 2} = b0 + b1\ result_{method\ 1}$$

The estimated values of the intercept and the slope are compared to the values expected for perfect concordance, i.e. an intercept of zero and a slope of one.

Turpeinen *et al.* used the Deming technique of linear regression in their comparison study. The

more familiar technique of ordinary linear regression analysis has as one of its underlying assumptions that the x variable has no appreciable measurement variability (Berry 1993). In order for this assumption to be satisfied, the variability in the x variable must be small compared to the variability in the y variable. In method comparison studies, however, it is typical for the x variable, i.e. the test results obtained using one method, to have a variability comparable to that of the y variable, the test results obtained using the other method. Consequently, ordinary regression analysis is usually not an appropriate regression technique for a method comparison study. Instead, one must use a linear regression technique that takes into account variability in the measurement of the x variable. One such "errors-in-variables" technique is the Deming method (Strike 1996). It has been found to be among the most reliable of the errors-in-variables linear regression techniques (Wakkers *et al.* 1975, Riggs *et al.* 1978, Linnet 1993). Weighted regression modifications of the Deming method are available for data sets in which the variance of the data pairs is not constant over the range of measurement (Riggs *et al.* 1978, Linnet 1990, 1993). The modification developed by Linnet applies to data sets in which the variance increases proportionally with analyte concentration.

Turpeinen *et al.* present graphs of the data and regression lines for each of the three method pairings. For the IMx and Diamat method pairing the graph is:



The authors found that the IMx and Diamat methods showed the best result concordance by (unweighted) Deming regression (note that the authors have been careless in their terminology, when they write "correlation" they mean "regression"),

> The correlation between IMx, calculated as %$HbA_{1c,}$ and the Bio-Rad Diamat (Fig 2A) gave the following results: IMx = 1.16 Diamat - 0.98 (r = 0.922). The good correlation is explained by the fact that the IMx assay has been standardized with an ionexchange HPLC method, with the Diamat assay as a secondary reference HPLC system maintained in close calibration to the primary reference HPLC assay.

The 95% confidence intervals on the estimates of the slope and intercept are 1.152 to 1.170 and -0.986 to -0.980, respectively. The confidence interval for the slope does not include 1 so the estimate is statistically different from 1. The confidence interval for the intercept does not include 0 so it is statistically different from 0. As neither parameter is equal to the value expected of perfect concordance, bias is present. When the intercept is not 0, the data are said to show constant bias. When the slope is not equal to 1, the data show proportional bias. When both parameters do not equal their ideal values, as here, the bias is referred to as mixed constant and proportional. The paper states

> the confidence limits for the slope and intercept values were calculated with the jackknife method

The jackknife method is a nonparametric technique for generating empirical likelihood distributions for the parameters of a statistical model (Mooney and Duval 1993). In this case, the statistical model is a line and the parameters are its slope and intercept. The jackknife method has been found to be a reliable way to calculate the parameter confidence intervals when either unweighted or weighted Deming regression is used (Linnet 1993). Parametric approaches for the calculation of the confidence intervals can be used if the methods have relatively constant variability over the clinical range of values for the analyte. Calculation of the exact confidence intervals is complex (Creasy 1956) but highly accurate, simple approximations are available (Strike 1996). The approximate confidence interval for the slope is,

$$b1 \pm t_c \text{ standard error of } b1$$

in which

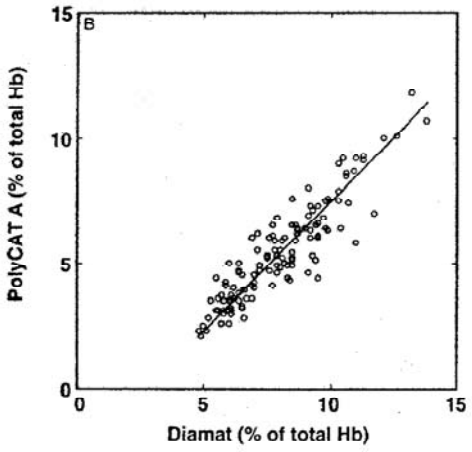
$$\text{standard error of } b1 = \sqrt{\frac{b1^2(1-r^2)/r^2}{n-2}}$$

where r^2 is the square of the correlation coefficient and n is the number of data pairs. The approximate confidence interval for the intercept is,

$$b0 \pm t_c \text{ standard error of } b1 \sqrt{\frac{\Sigma x^2}{n}}$$

where Σx^2 is the sum of the squared x values.

The graph of the data and Deming regression line for the Diamat and PolyCAT A pairing is:



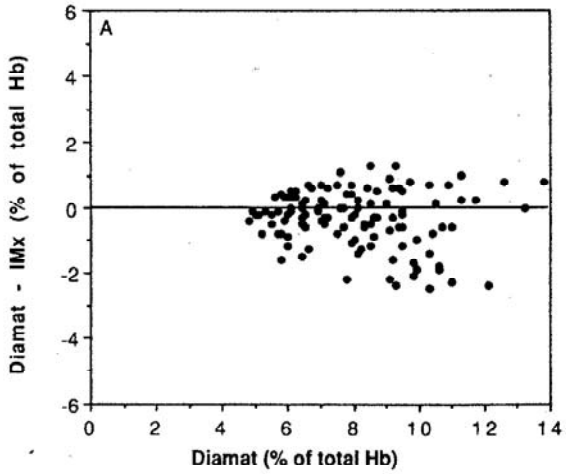
The result concordance of these methods is much less good that that between the Diamat and IMX methods:

> When the two ion-exchange chromatographic methods, PolyCAT A and Diamat, were compared (Fig 2B) . . . the regression equation (PolyCAT A = 1.03 Diamat - 2.84) shows that much lower results were obtained by ion-exchange chromatography with high resolution.

The 95% confidence intervals on the estimates of the slope and intercept are 1.026 to 1.038 and -2.842 to -2.834, respectively. Although the slope is statistically different from 1, the difference is small, so the authors conclude that the lack of concordance appears to be due largely to constant bias. The authors offer the following explanation for the bias:

> This might be due to the fact that the Diamat method also measures carbamylated and acetylated forms of Hb … and possibly some other derivatives formed in blood during storage, which can be separated from the HbA_{1c} peak by using methods with higher resolution. Our PolyCAT A assay has been optimized to separate different Hb variants from HbA_{1c}. This apparently partly explains the lower results obtained by this method. However, the difference between Diamat method and PolyCAT A assay cannot be explained only by carbamylated and acetylated Hbs, for which concentrations $< 0.4\%$ have been reported …

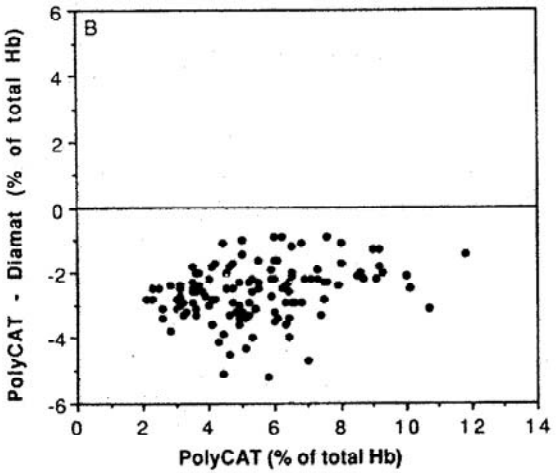
Difference analysis. Difference analysis was introduced as an approach for evaluating method comparisons by Altman and Bland (1983). This approach strives to avoid the shortcomings of ordinary regression analysis for method comparison by directly measuring how well each result pair agrees in terms of the difference in the values of the results. The differences are plotted against the average values of the result pairs (Bland and Altman 1986, Hollis 1996) or, as in the article by Turpeinen *et al.,* the differences are plotted against the results of one of the methods. In the comparison of the IMx and Diamat methods, the plot, called a difference plot, is:



Examination of the difference plot suggests that there is a small inter-method bias present because there are more negative differences than positive differences. If there were no bias present, the differences would be evenly distributed about the line of zero difference. The authors describe the pattern as follows:

> The bias observed (Fig 3A) suggests that the IMx method gives slightly higher results than the Diamat method at high amounts but similar results at normal amounts of HbA_{1c}.

The difference plot for the Diamat and PolyCAT A method pairing is:



Here the bias is clear. The authors comment:

> The differential plots show that the negative bias of ~ 2-3% of total BB is seen at all values of HbA_{1c} when PolyCAT A is compared with Diamat

Assessment of clinical equivalence

As stated earlier, clinical equivalence of two analytical methods means that they can be used interchangeably. In practical terms, clinical equivalence means two things: that the results of the two methods show a high degree of concordance and that the reference ranges for the measured analyte, as determined using the two methods, are essentially identical. The degree of concordance and the closeness of the agreement of the reference ranges that are required in order to consider two method clinically equivalent are matters of clinical judgment which may be codified in recommendations promulgated by professional societies or in standards imposed by regulatory agencies. Statistical evidence is important in coming to this decision but it is not the only consideration. For instance, the confidence interval for the estimate of the intercept of the regression line for paired results may indicate that it is statistically different from zero. This indicates the presence of a bias in the methods. However, the magnitude of the bias may be considered to be clinically insignificant and the results of the methods deemed to be highly concordant.

Figure 2.9 demonstrates a graphical approach for judging if two methods are clinical equivalent in terms of the degree of concordance of the methods. In this figure the data pairs have been plotted as they

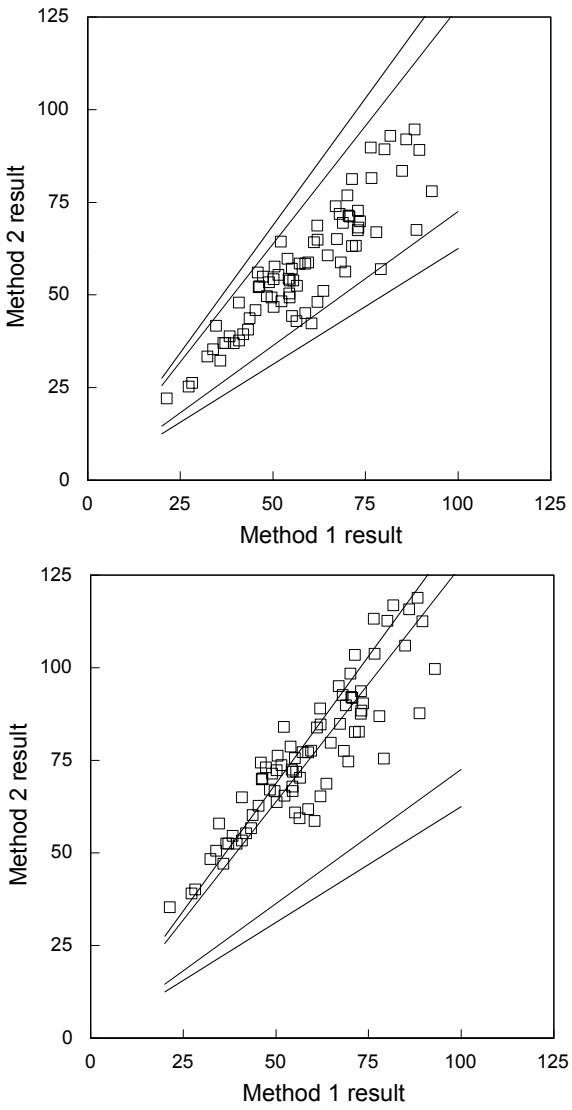


Figure 2.9 Two sets of hypothetical method comparison data. The data are shown as symbols. Clinical equivalence boundary lines are indicated.

would be for a regression analysis. The lighter, inner lines demarcate the region in which 95% of the data pairs will be if the methods are, on average, perfectly concordant. Because each sample is only measured once by each method in the usual method comparison, individual method measurement variability results in less than perfect concordance even if the methods are, on average, perfectly concordant. If each sample were measured repeatedly by each method, the average result values obtained by two methods that were, on average, perfectly concordant would show exact agreement. These boundary lines are constructed using the formula,

$$range = analyte\ concentration \pm z_c\ SD_{result\ pairs}$$

where, for the 95% range, z_c equals 1.96, and

$$SD_{result\ pairs} = \sqrt{var_{method\ 1} + var_{method\ 2}}$$

The width of the range will vary with analyte concentration if the variance of either method depends upon the concentration. In the figure, the variance was treated as being proportional to analyte concentration. The darker, outer lines delimit the region in which 95% of the data pairs will be if the methods are concordant to within a clinically acceptable amount of bias. These are the clinical equivalence boundary lines. These boundary lines are constructed using the formula,

$$range = analyte\ concentration \pm (z_c\ SD_{result\ pairs} + acceptable\ bias)$$

The upper graph in Figure 2.9 shows the hypothetical results of a comparison of two highly concordant methods. All of the data points are inside of the clinical equivalence boundary lines (here the bias criterion is a relative bias of less than 7.5%). Using the approximate formula (Newcombe 1998),

$$confidence\ interval = \frac{estimate + \frac{z_c^2}{2N} \pm z_c \sqrt{\frac{estimate(1 - estimate)}{N} + \frac{z_c^2}{4N}}}{1 + \frac{z_c^2}{N}}$$

where N is the number of data pairs, the approximate 95% confidence interval on the proportion is 95.4 to 100%. Because it is statistically certain that at least 95% of the data pairs are contained within the boundary lines, the concordance of the methods satisfies the criterion for clinical equivalence. The lower graph shows hypothetical results from two methods which are discordant. Seventy percent of the data points are inside of the clinical equivalence boundary lines with a 95% confidence interval of 51.5 to 72.3%. It is, therefore, statistically certain that fewer than 95% of the data pairs are contained within the boundary lines. Hence, the methods are not clinically equivalent due, in this case, to the presence of an unacceptably large inter-method bias.

A similar graphical approach can be taken when the result data are plotted as differences (Petersen *et al.* 1997). This is shown in Figure 2.10 for the same hypothetical data that were used to make Figure 2.9. Notice that in this figure, the x-axis is the result as measured by the field method already in place (method 1). When plotted in this fashion, the

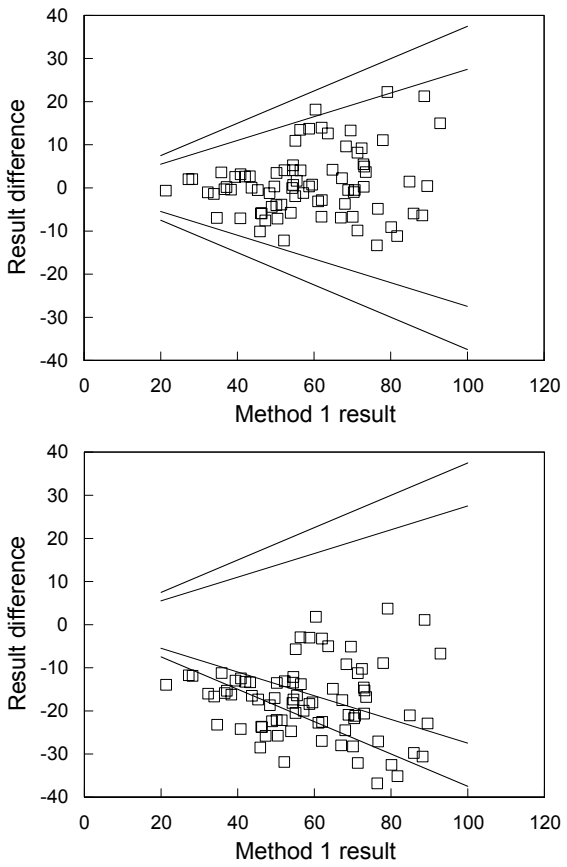


Figure 2.10 The same hypothetical method comparison data as in Figure 2.9 but here presented as difference plots. The data are shown as symbols. Clinical equivalence boundary lines are indicated.

graphs of the two data sets reveal exactly the same percentage of data points inside of the clinical equivalence boundary lines.

Based upon the difference analysis findings, Turpeinen *et al.* conclude that the methods they studied do not demonstrate a clinically acceptable degree of concordance. The authors also constructed reference ranges for HbA_{1c} using two of the methods under study:

> Reference values for the Diamat and PolyCAT A methods were determined by using 60 freshly drawn blood samples from healthy controls

They found that the ranges are not similar:

> For IMx a reference range of 4.5–5.5% has been reported (7). Our estimates [of] the reference values for HbA_{1c} with the PolyCAT A and Diamat methods with samples from 60

healthy controls were, for Diamat, mean (±SD) 5.13% ± 0.33%, the reference range (mean±2SD) 4.5–5.8%, and the total range 4.4–6.1%. For the PolyCAT A method the mean (±SD) was 3.43% ± 0.47%, the reference range 2.5–4.4%, and the range 2.6–5.0%.

Because the methods do not produce identical reference ranges and because they are not highly concordant, the authors decided that the methods are not clinically equivalent.

REFERENCES

Altman DG and Bland JM. 1983. Measurement in medicine: the analysis of method comparison studies. Statistician 32:307.

Berry WD. 1993. *Understanding Regression Assumptions*. Sage Publications, Newbury Park CA.

Bezeau M and Endrenyi L. 1986. Design of experiments for the precise estimation of dose-response parameters: the Hill equation. J theor Biol 123:415.

Bishop J and Nix ABJ. 1993. Comparison of quality-control rules used in clinical chemistry laboratories. Clin Chem 39:1638.

Bland JM and AltmanDG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i:307.

Bookbinder MJ and Panosian KJ. 1986. Correct and incorrect estimation of within-day and between-day variation. Clin Chem 32:1734.

Box GEP and Draper NR. 1987. *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, New York.

Box GEP, Hunter WG, and Hunter JS. 1978. *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons, New York.

Büttner J, Borth R, Boutwell HJ, Broughton PMG, and Bowyer RC. 1980a. Approved recommendation (1978) on quality control in clinical chemistry. Part 1. General principles and terminology. J Clin Chem Clin Biochem 18:69.

Büttner J, Borth R, Boutwell HJ, Broughton PMG, and Bowyer RC. 1980b. Approved recommendation (1978) on quality control in clinical chemistry. Part 2. Assessment of analytical methods for routine use. J Clin Chem Clin Biochem 18:78.

Büttner J, Borth R, Boutwell HJ, Broughton PMG, and Bowyer RC. 1980c. Approved recommendation (1978) on quality control in clinical chemistry. Part 3. Calibration and control materials. J Clin Chem Clin Biochem 18:855.

Büttner J, Borth R, Boutwell HJ, Broughton PMG, and Bowyer RC. 1983a. Approved recommendation (1978) on quality control in clinical chemistry. Part 4. Internal quality control. J Clin Chem Clin Biochem 18:877.

Büttner J, Borth R, Boutwell HJ, Broughton PMG, and Bowyer RC. 1983b. Approved recommendation (1978) on quality control in clinical chemistry. Part 5. External quality control. J Clin Chem Clin Biochem 18:885.

Cotlove E, Harris EK, and Williams GZ. 1970. Biological and analytical components of variations in long-term studies of serum constituents in normal subjects. III. Physiological and medical implications. Clin Chem 16:1028.

Creasy MA. 1956. Confidence limits for the gradient in the linear functional relationship. J Roy Stat Soc B 18:65.

Dybkaer R. 1995. Result, error and uncertainty. Scand J Clin Lab Invest 55:97.

Dybkaer R. 1997. Vocabulary for use in measurement procedures and description of reference materials in laboratory medicine. Eur J Clin Chem Clin Biochem 35:141.

Ekins RP. 1983. The precision profile: its use in assay design, assessment and quality control. In Hunter WM, and Corrie JET (eds). *Immunoassays for Clinical Chemistry*. Second edition. Churchill Livingston, Edinburgh.

Ekins R and Edwards P. 1997. On the meaning of "sensitivity". Clin Chem 43:1824.

Emancipator K and Kroll MH. 1993. A quantitative measure of nonlinearity. Clin Chem 39:766.

Fedorov VV. 1972. *Theory of Optimal Experiments*. Academic Press, New York.

Fraser CG, Petersen PH, Libeer J-C, and Ricos C. 1997. Proposals for setting generally applicable quality goals solely based on biology. Ann Clin Biochem 34:8.

Gardner E. 1985. Exponential smoothing: the state of the art. J Forecast 4:1.

Gautschi K, Keller B, Keller H, Pei P, and Vonderschmitt DJ. 1993. A new look at the limits of detection (L_D), quantification (LQ) and power of definition (PD). Eur J Clin Chem Clin Biochem 31:433.

Gowans EMS, Hyltoft Petersen P, Blaabjerg O, and Hřrder M. 1988. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 48:757.

Gowans EMS, Hyltoft Petersen P, Blaabjerg O, and Hřrder M. 1989. Analytical goals for the estimation of non-Gaussian reference intervals. Scand J Clin Lab Invest 49:727.

Harris EK. 1979. Statistical principles underlying analytic goal-setting in clinical chemistry. Am J Clin Pathol 72:374.

Hollis S. 1996. Analysis of method comparison studies. Ann Clin Biochem 33:1.

Keller H and Passing H. 1989. Performance profiles: new tools for characterization and comparison of clinical chemical results. J Clin Chem Clin Biochem 27:613.

Kroll MH and Chesler R. 1992. Rationale for using multiple regression analysis with complex interferences. Eur J Clin Chem Clin Biochem 30:415.

Kroll MH and Elin RJ. 1994. Interference with clinical laboratory analyses. Clin Chem 40:1996.

Kroll MH and Emancipator K. 1993. A theoretical evaluation of linearity. Clin Chem 39:405.

Linnet K. 1990. Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med 9:1463.

Linnet K. 1993. Evaluation of regression procedures for methods comparison studies. Clin Chem 39:424.

Luque de Castro MD, Quiles R, Fernández-Romero JM, and Fernández E. 1995. Continuous-flow assay with immobilized enzymes for determining of inorganic phossphate in serum. Clin Chem 41:99.

Motulsky HJ and Ransnas LA. 1987. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. FASEB J 1:365.

Mooney CZ and Duval RD. 1993. *Bootstrapping. A Nonparametric Approach to Statistical Inference*. Sage Publications, Newbury Park CA.

Newcombe RG. 1998. Two-sided confidence intervals for the single proportion: comparisons of seven methods. Stat Med 17:857.

Nix ABJ, Rowlands RJ, Kemp KW, WIlson DW, and Griffiths K. 1987. Internal quality control in clinical chemistry: a teaching review. Stat Med 6:425.

Pardue HL. 1997. The inseparable triad: analytical sensitivity, measurement uncertainty, and quantitative resolution. Clin Chem 43:1831.

Parvin CA. 1992. Comparing the power of quality-control rules to detect persistent systemic error. Clin Chem 38:358.

Passey RB and Maluf KC. Linearity and calibration. A clinical laboratory perspective. Arch Pathol Lab Med 116:757.

Petersen PH, Ricós C, Stöckl D, Libeer JC, Baadenhuijsen H, Fraser C, and Thienpont L. 1996. Proposed guidelines for the internal quality control of analytical results in the medical laboratory. Eur J Clin Chem Clin Biochem 34:983.

Petersen PH, Stöckl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, Lassen JF, and Kjeldsen J. 1997. Graphical interpretation of a field method with a reference method by use of difference plots. Clin Chem 43:11.

Riggs DS, Guarnieri JA, and Addelman S. 1978. Fitting straight lines when both variables are subject to error. Life Sciences 22:1305.

Rodbard D. 1978. Statistical estimation of the minimal detectable concentration ("sensitivity") for radioligand assays. Anal Biochem 90:1.

Sadler WA, Murray LM, and Turner JG. 1992. Minimum distinguishable difference in concentration: a clinically oriented translation of assay precision summaries. Clin Chem 38:1773.

Sadler WA and Smith. 1990. Use and abuse of imprecision profiles: some pitfalls illustrated by computing and plotting confidence intervals. Clin Chem 36:1346.

Sadler WA, Smith MH, and Legge HM. 1988. A method for direct estimation of imprecision profiles, with reference to immunoassay data. Clin Chem 34:1058.

Sebastián-Gámbaro MÁ, Lirón-Hernandez FJ, and Fuentes-Arderiu X. 1997. Intra- and inter-individual biological variability data bank. Eur J Clin Chem Clin Biochem 35:845.

Smith FA and Kroft SH. 1997. Optimal procedures for detecting analytic bias using patient samples. Am J Clin Pathol 108:254.

Steinberg DM and Hunter WG. 1984. Experimental design: review and comment. Technometrics 26:71.

Stöckl D. 1996. Metrology and analysis in laboratory medicine: a criticism from the workbench. Scand J Clin Lab Invest 56:193.

Stöckl D, Baadenhuijsen H, Fraser CG, Libeer J-C, Petersen PH, and Ricós C. 1995. Desirable routine analytical goals for quantities assayed in serum. Eur J Clin Chem Clin Biochem 33:157.

Strike PW. 1996. *Measurement in Laboratory Medicine: A Primer on Control and Interpretation*. Butterworth-Heinemann, Oxford.

Tholen DW. 1992. Alternative statistical techniques to evaluate linearity. Arch Pathol Lab Med 116:746.

Turpeinen U, Karjalainen U, and Stenman U-H. 1995. Three assays for glycohemoglobin compared. Clin Chem 41:191.

Wakkers PJM, Hellendoorn HBA, Op de Weegh GJ, and Heerspink W. 1975. Applications of statistics in clinical chemistry. A critical evaluation of regression lines. Clin Chim Acta 64:173.

Westgard JO, Barry PL, Hunt MR, and Groth T. 1981. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 27:493.