Clinical Chemistry 45:6
882–894 (1999)

Laboratory Management

Necessary Sample Size for Method Comparison Studies Based on Regression Analysis

Kristian Linnet

Background: **In method comparison studies, it is of importance to assure that the presence of a difference of medical importance is detected. For a given difference, the necessary number of samples depends on the range of values and the analytical standard deviations of the methods involved. For typical examples, the present study evaluates the statistical power of least-squares and Deming regression analyses applied to the method comparison data.**

Methods: **Theoretical calculations and simulations were used to consider the statistical power for detection of slope deviations from unity and intercept deviations from zero. For situations with proportional analytical standard deviations, weighted forms of regression analysis were evaluated.**

Results: **In general, sample sizes of 40–100 samples conventionally used in method comparison studies often must be reconsidered. A main factor is the range of values, which should be as wide as possible for the given analyte. For a range ratio (maximum value divided by minimum value) of 2, 544 samples are required to detect one standardized slope deviation; the number of required samples decreases to 64 at a range ratio of 10 (proportional analytical error). For electrolytes having very narrow ranges of values, very large sample sizes usually are necessary. In case of proportional analytical error, application of a weighted approach is important to assure an efficient analysis; e.g., for a range ratio of 10, the weighted approach reduces the requirement of samples by >50%.**

Conclusions: **Estimation of the necessary sample size for a method comparison study assures a valid result; either no difference is found or the existence of a relevant difference is confirmed.**

A common task in the laboratory is to compare a new method with an established one to assess whether the new measurements are comparable with the existing ones. The question may be whether the new measurements are interchangeable with the existing ones from a clinical point of view, or whether proficiency testing demands are likely to be fulfilled. Various sources may be consulted to evaluate the differences that are considered of importance. Medically significant differences have been assessed on the basis of clinicians' points of view *(1)*, biological variation *(2)*, or combinations of these principles *(3)*. For some analytes, e.g., cholesterol and other lipids, organizations have recommended specific analytical goals that take into account the medical use of these analytes *(4, 5)*. In the context of external quality assessment, regulatory authorities have decided on analytical tolerance limits that should be achieved, e.g., CLIA 88 demands *(6)*. On the basis of these kinds of sources, one may reach a conclusion concerning relevant critical differences that should be detected at one or more decision levels. A typical decision level might be the upper edge of the 95% reference interval. Other levels might be dictated by medical intervention limits, e.g., in the context of serum cholesterol concentrations.

The rational design of a method comparison study should take into account the relevant critical differences that should be detected at selected decision levels. Commonly, the measurements of two analytical methods are compared by a regression analysis procedure, which allows the detection of a possible constant systematic difference (intercept deviation from zero) and a proportional systematic difference (slope deviation from unity). The investigator should then consider whether the study design is likely to disclose these critical differences. Important factors in this context are the range of measurements, the analytical standard deviations (SD_as)[1] of the

Laboratory of Clinical Biochemistry, Psychiatric University Hospital, Skovagervej 2, DK-8240 Risskov, Denmark. Fax 45 86170778; e-mail linnet@post7.tele.dk.

Received September 11, 1998; accepted March 25, 1999.

[1] Nonstandard abbreviations: SD_a, analytical standard deviation; OLR, ordinary least-squares regression analysis; WLR, weighted least-squares regression analysis; *D*, estimated difference; Δ, true difference; and CV_a, analytical coefficient of variation.

involved methods, and the number of samples. These factors determine the statistical power of a method comparison study, i.e., the ability of the data analysis procedure to verify the presence of a given systematic difference. In this study, some prototype situations in clinical chemistry are evaluated, and guidelines concerning necessary sample sizes are tabulated for typical cases. In situations with constant SD_as, unweighted regression procedures are considered, i.e., ordinary least-squares regression analysis (OLR) and Deming regression analysis. For cases involving SD_as that are proportional to the measurement level, the corresponding weighted regression procedures are primarily taken into account. In the *Appendix*, specific formulas are presented for the relationship between statistical power and sample size in regression analysis.

Materials and Methods

METHOD COMPARISON MODEL

Taking into account that an analytical method measures analyte concentrations with some uncertainty, one must distinguish between the measured value (x_i) and the target value (X_i) of a sample subjected to analysis by a given method. The latter is the mean result we would obtain if the given sample was measured an infinite number of times. The measured value is likely to deviate from the target value by some small random amount (ϵ or δ). For a given sample measured by two analytical methods, we have:

$$x_i = X_i + \epsilon_i$$

$$y_i = Y_i + \delta_i$$

The dispersion of measured values around the target value depends on the SD_a of the method. A linear relationship between the target values of the two methods is assumed:

$$Y_i = \alpha_0 + \beta X_i$$

To estimate α_0 and β correctly, a regression procedure taking errors in both x and y into account is preferable, e.g., the Deming method *(7–10)*. In this procedure, the sum of squared distances from measured sets of values (x_i, y_i) to the regression line is minimized at an angle determined by the ratio between the SD_as of x and y. In Fig. 1A, the symmetric case is illustrated with a regression slope of one and equal SD_as for x and y. The most widely used regression procedure in method comparison studies, OLR, does not take errors in x into account and thus provides a downward biased slope estimate *(7)*. In situations with a wide range of x values, this bias may be negligible and OLR may be used for estimation of slope and intercept. In OLR, the sum of squared distances from (x_i, y_i) to the line is minimized in the vertical direction (Fig. 1B).

Another methodological problem concerns the question of whether the SD_as are constant. For most clinical chemical compounds, the SD_as vary with the measurement level *(11)*. In cases with a considerable range, i.e., a decade or more, this phenomenon should also be taken into account in the regression analysis. The Deming method should then be carried out as a weighted analysis, e.g., assuming proportional SD_as *(12)*. In the weighted modification of the Deming procedure, distances from (x_i, y_i) to the line are inversely weighted according to the

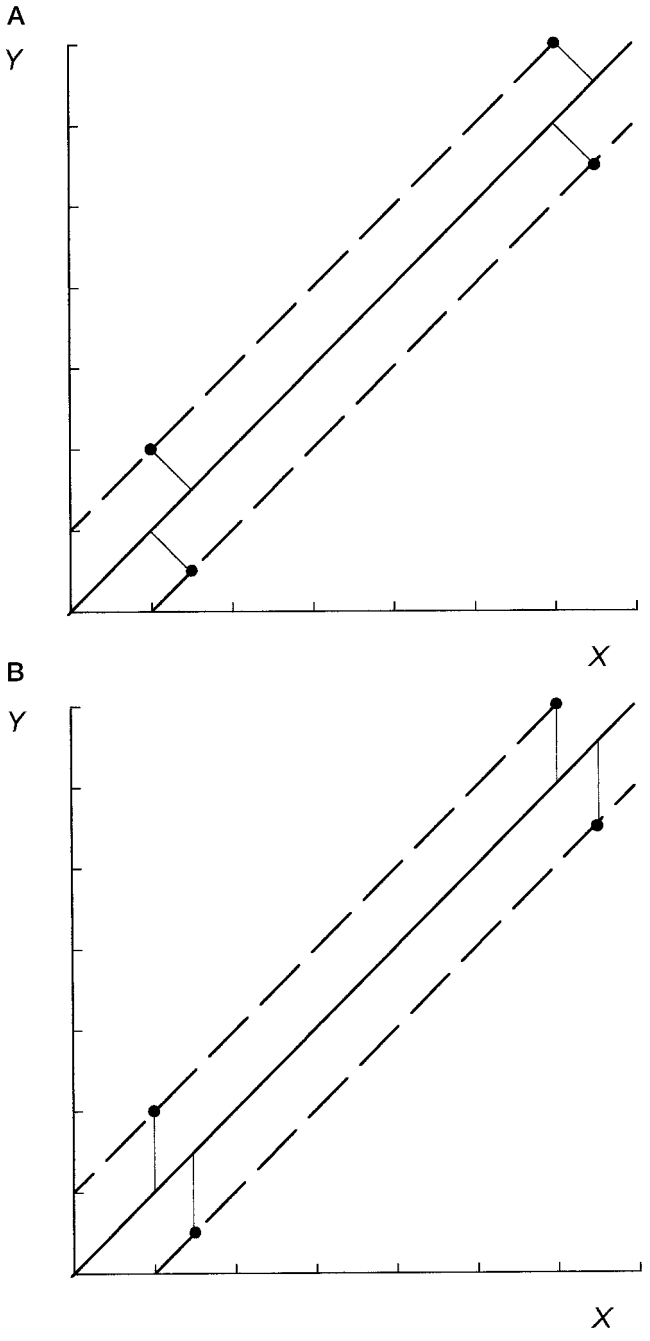

Fig. 1. Schematic outline of distances from data points to the line in Deming regression (*A*) and outline of vertical distances from data points to the line in OLR (*B*).

In *panel A*, SD_{ax} and SD_{ay} are equal, and the slope is one.

squared SD_as at a given level (Fig. 2A). In the same way, least-squares regression analysis may also be carried out in a weighted modification (WLR), in which the distances from (x_i, y_i) to the line in the vertical direction are inversely weighted according to the squared SD_a value (Fig. 2B) *(9, 13)*. The regression procedures, which are outlined in the *Appendix*, were performed using the program CBstat developed by the author. The program runs under Windows 95/98/NT.

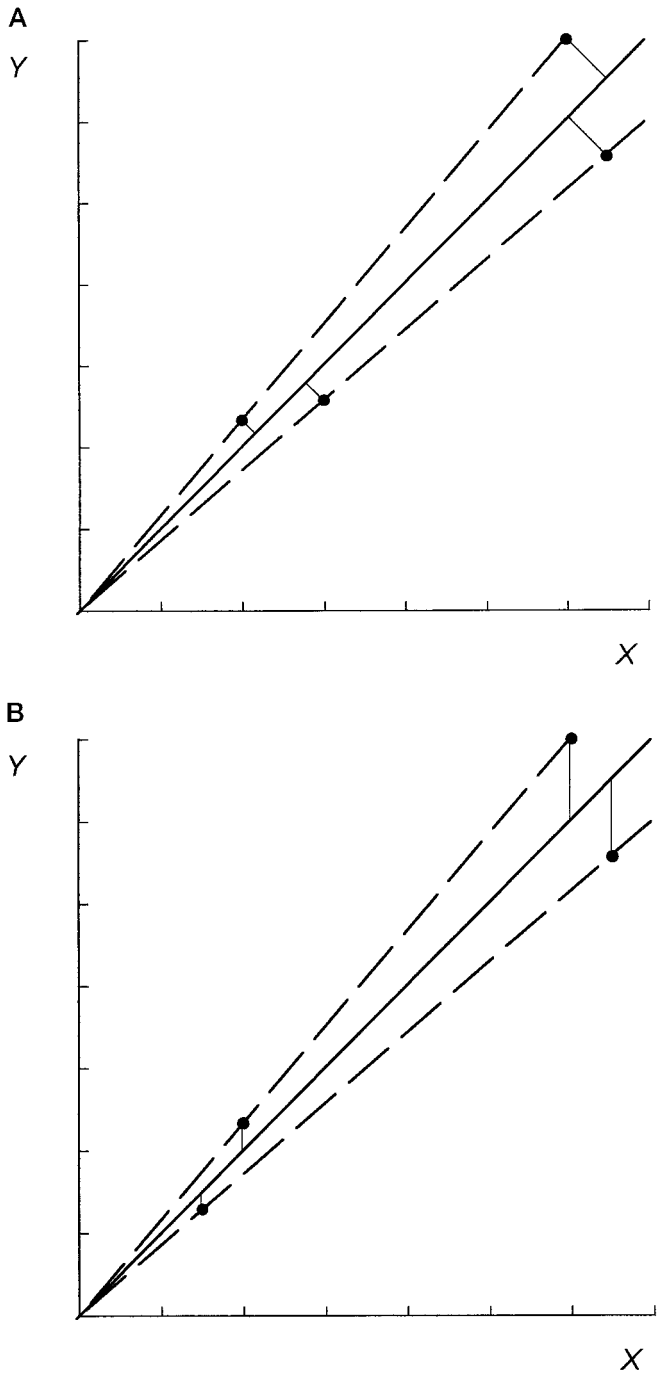

Fig. 2. Distances from data points to the line in weighted Deming regression assuming proportional SD_as (*A*) and vertical distances from data points in WLR (*B*).

In *panel A*, SD_{ax} and SD_{ay} are assumed equal.

DETECTION OF SYSTEMATIC DIFFERENCES BETWEEN METHODS

A systematic difference between two methods is identified if the estimated intercept differs significantly from zero or if the slope deviates significantly from 1. This is decided on the basis of *t*-tests:

$$t = (a_0 - 0)/\mathrm{SE}(a_0)$$

$$t = (b - 1)/\mathrm{SE}(b)$$

$\mathrm{SE}(a_0)$ and $\mathrm{SE}(b)$ are the standard errors of the estimated intercept a_0 and slope b, respectively. For OLR and WLR, the standard errors are calculated from formulas; for the Deming and weighted Deming procedures, one may apply a computerized resampling principle called the jackknife procedure (*Appendix*) *(12–15)*. Notice that Latin letter symbols, e.g., a_0 and b, denote sample estimates, whereas Greek letters are used for true, population values (α_0 and β in this case).

Having estimated a_0 and b, it is possible to estimate the systematic difference between the methods, D_c, at a selected level X_c (Fig. 3):

$$D_c = Y_{estc} - X_c = a_0 + (b - 1)X_c$$

Y_{estc} is the estimated Y value at X_c. Notice that D_c refers to the systematic difference, i.e., the difference between target values, and so it is not a total error including random measurement errors.

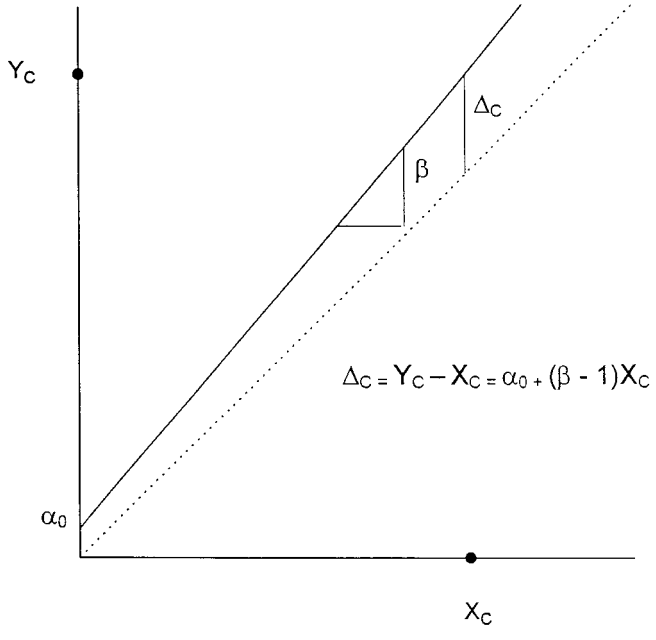

Fig. 3. Illustration of the systematic difference Δ_c between two methods at a given level X_c according to the regression line.

The difference is a result of a constant systematic difference (intercept deviation from zero) and a proportional systematic difference (slope deviation from unity). The *dotted line* represents the diagonal $Y = X$.

In the planning phase of a method comparison study, one should consider the size of the medically significant difference or critical difference, Δ_c, that should be detected at a given level X_c. One may then derive the needs for detecting critical sizes of α_0 and slope deviation from unity ($\beta - 1$):

$$\Delta_c = Y_c - X_c = \alpha_0 + (\beta - 1)X_c$$

Evaluations of slope and intercept deviations are presented in the following sections.

STATISTICAL POWER CONSIDERATIONS IN METHOD COMPARISON STUDIES

Having decided from above the values of α_0 and ($\beta - 1$) that should be detected, the next step is to design the method comparison study appropriately. A relevant range of values should be included, i.e., covering the range of clinical interest. A uniform distribution of values over the range is preferable and is primarily supposed in the present evaluation. We next must consider the SD_as of the methods. Two situations should be considered: the presence of constant and nonconstant SD_as. The first possibility is of interest mainly for measurements involving a narrow range of values, e.g., in electrolyte methods. When ranges cover one or more decades, it is important to take into account that the SD_a usually varies with the measurement level *(11)*. Quite often a proportional relationship approximately applies in clinical chemistry, implicating that the analytical coefficient of variation (CV_a) is approximately constant over the measurement range. Finally, specific values for the SD_as or CV_as of the methods should be assigned from available quality-control data. According to the formula presented below and explained in more detail in the *Appendix*, we now have the necessary information to plan the comparison study, i.e., to decide on the necessary number of samples.

A general, simplified formula for the approximation of the necessary sample size for detection of a difference Δ with regard to slope deviation from unity or intercept deviation from zero *(16)* is:

$$N = (c/\Delta)^2(t_{p/2} + t_{1-q})^2$$

where c is a constant determining the standard error ($c/\sqrt{N}$) of the estimated difference D, which corresponds to the true difference Δ. $t_{p/2}$ depends on the significance level P (type I error) and is 1.96 (asymptotically) for $P \leq 5\%$. t_{1-q} reflects the statistical power ($1 - q$), which is the probability of verifying a real difference Δ. The complement to the power is the type II error (q), which is the probability of overlooking a real difference Δ. For a traditional power level of 90%, t_{1-q} takes the value 1.28. Finally, the sample size is in principle inversely related to the squared difference Δ, i.e., if a given difference is halved, the sample size requirement is increased by a factor four. At small-to-moderate sample sizes, however, adjustment of the t value from the asymptotic value and the general impact of approximations disturb this relationship somewhat (*Appendix*).

The relationship between the null hypothesis situation of no difference and the alternative hypothesis of the presence of a real difference Δ is depicted schematically in Fig. 4, which outlines the hypothetical situation corresponding to a set of repeated method comparison studies that yield observed differences D that are distributed around the true difference, which is zero under the null hypothesis of no difference and equal to Δ under the alternative hypothesis. The larger the sample size is, the more narrow the dispersions of observed differences around the true values are. Thus, for a given Δ and type I error, the power increases with the sample size.

Results

The necessary sample sizes for a series of standard method comparison situations in clinical chemistry have been tabulated in Tables 1 and 2. A type I error (significance level) of 5% and a power of 90% have been supposed. Table 1 concerns the situation with constant SD_as over the measurement range, and Table 2 covers cases with proportional SD_as.

CONSTANT SD_aS

Table 1 covers intervals with ratios from 1.25 to 10 for the maximum value divided by the minimum value (range ratio = maximum value/minimum value). The other

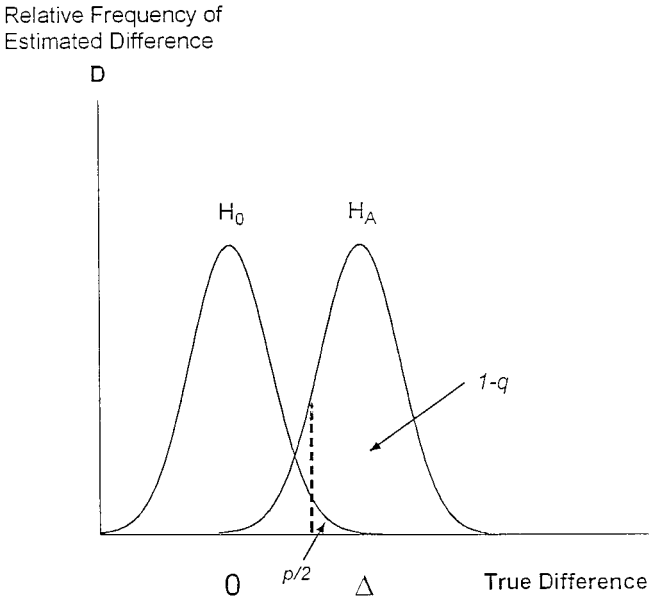

Fig. 4. Schematic illustration of distributions of differences D under the null hypothesis (H_0) of no real difference and the alternative hypothesis (H_A) of the presence of a real difference Δ.

The *vertical dotted line* indicates the limit of statistical significance. p is the type I error (5%), and $1 - q$ is the power (90%).

Table 1. Sample size table for comparison of methods with constant SD_as using Deming regression analysis.[a]

	Range ratios								
	1.25	**1.5**	**2**	**2.5**	**3**	**4**	**5**	**8**	**10**
$\Delta\beta_{st} = (\beta - 1)/CV_a$[b]									
1	5104	1575	567	343	256	182	150	116	108
2	1276	410	152	90	69	48	39	32	27
3	585	185	70	42	32	25	20	16	15
4	325	104	41	27	20	15	13	11	≤10
$\Delta\alpha_{Ost} = (\alpha_0 - 0)/SD_a$									
1	5125	1596	588	364	273	196	168	125	120
2	1281	410	152	91	71	52	44	35	32
3	580	187	71	42	33	26	22	17	16
4	330	110	40	27	21	15	13	11	≤10

[a] Necessary sample sizes for test of slope deviation from 1 or intercept deviation from zero by Deming regression analysis given constant SD_as: $SD_{ax} = SD_{ay} = SD_a$. Uniform *x* and *y* distributions on intervals with the given range ratio. The range ratio is the maximum value divided by the minimum value of the considered interval. OLR (given $SD_{ax} = 0$) requires one-half as many samples as sample sizes listed. Type I error, 5%; power, 90%.

[b] CV_a refers to the CV at the middle of the given interval (SD_a/x_m).

entry in Table 1 is the standardized Δ value for slope or intercept. With regard to the slope, this value refers to the slope deviation from unity measured in CV_a units:

$$\Delta\beta_{st} = (\beta - 1)/CV_a$$

Notice that although the SD_a is constant, the CV_a (expressed as a fraction) enters the formula, implying that the CV_a is not constant over the measurement range. The CV_a in the formula refers to the specific CV_a value at the middle of the interval of interest, i.e., $CV_a = SD_a/x_m$, where x_m is the mean of the interval for the analytes (*Appendix*).

With regard to the intercept deviation from zero, the standardized Δ value is:

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/SD_a$$

The standard situations presuppose that the analytical methods have identical SD_as and that the analyte values are uniformly distributed over the intervals of interest. Notice that if duplicate measurements are carried out by each method, the SD_a value for single measurements should be reduced by a factor of $\sqrt{2}$. Table 1 has only selected entry values. With regard to standardized deviations that are not covered, an approximate sample size may be obtained by inter- or extrapolation. At large sample sizes, squared inter- or extrapolation is reasonable, but for small sample sizes, this relationship is not exactly valid. Approximate interpolations can also be carried out for the tabulated range ratios. Given assumptions not covered in Table 1, estimation of *N* may be performed on the basis of formulas described in the *Appendix*. Moreover, one should consider adding some additional samples to take nonideal conditions into account, such as target value distributions that are not exactly uniform over the given interval (discussed later in the text). The tabulated values refer to application of Deming regression analysis. If OLR is applied, the required sample sizes are one-half the tabulated values. However, to apply OLR correctly, the *x* measurements should be without random measurement errors.

It is apparent from Table 1 that the range of values is very important with regard to the required sample size for detection of a given standardized slope or intercept deviation. For the very narrow ranges characteristic of electrolyte measurement methods, detection of a standardized slope or intercept deviation equal to one may require >1000 samples. On the other hand, the sample size requirements may be rather modest for analytes with values dispersed over wider ranges.

In the next section, Table 1 is used for evaluation of sample size requirements for a method comparison study of two electrolyte methods, i.e., a situation with a small range ratio and constant SD_as.

Planning a Comparison of Two Potassium Methods (Constant SD_as)

We first decide on the critical differences that should be detected. For convenience, we take as a basis the CLIA 88 rule of 0.5 mmol/L as the acceptable error throughout the measurement range. Notice that CLIA 88 rules relate to the total error in relation to a target value of a quality-control sample:

$$Total\ Error = Systematic\ Error + 1.65\ SD_a$$

The factor 1.65 corresponds to a total error that assumes that 95% of the measurements are within the given limit.

We consider here decision levels of 3 and 6 mmol/L and suppose in the present example that the SD_a is 0.09 mmol/L, which corresponds to a CV_a of 2% at the mean (4.5 mmol/L) of the considered range. Thus, the systematic difference that should be detected is:

$$\Delta_c = 0.5\ mmol/L - 1.65 \cdot 0.09\ mmol/L$$

$$= 0.35\ mmol/L$$

From the general formula:

$$\Delta_c = Y_c - X_c = \alpha_0 + (\beta - 1)X_c$$

we obtain at X_c = 3 mmol/L:

$$0.35 = 3.35 - 3 = \alpha_0 + (\beta - 1)3$$

This corresponds to a need for detecting α_0 equal to ± 0.35 mmol/L, if the systematic difference is ascribed to an intercept deviation. Relating the systematic difference to a slope deviation corresponds to a demand of detecting β = 1.12 (3.35/3), or 0.88. Similarly, at the upper decision level of X_c = 6 mmol/L, we have again the limits ± 0.35 mmol/L for detection of α_0, but now the demand for detecting a slope deviation has been sharpened to β = 1.06 (6.35/6) or 0.94.

Let us now consider the various factors in the estimation of sample size. The first factor to consider is the measurement range of 3–6 mmol/L, i.e., a range ratio of 2. We suppose, as mentioned above, that both methods have constant SD_as corresponding to a CV_a of 2%, i.e., 0.09 mmol/L, at the middle of the range. If duplicate sets of measurements are taken, the SD_a is reduced by a factor of $\sqrt{2}$, to 0.06 mmol/L. If we assume duplicate sets of measurements, the CV_a at the middle of the interval is 0.014. We are now able to convert the slope Δ value to a standardized value:

$$\Delta\beta_{st} = (\beta - 1)/CV_a = 0.06/0.014 = 4.3$$

The next factor to consider is the regression procedure, in this case, Deming regression analysis with a significance level (type I error) of 5% and a statistical power of 90%. To get the necessary sample size, we consult Table 1 and look under a range ratio of 2 and a standardized slope deviation of 4 and find the sample size, N = 41. When we use squared extrapolation, the sample size becomes 36 [$41 \times (4/4.3)^2$]. Notice that this value refers to 36 samples measured in duplicate with each method. If only single measurements are to be performed, the required number of samples would be doubled to 72.

Table 1 also covers cases studied by the use of OLR. Under the given assumptions, the approximate sample size requirement for OLR is obtained by dividing the numbers by two, i.e., N = 18 in this case (see *Appendix*). However, a correct statistical analysis based on OLR requires that the x method is without analytical errors (SD_{ax} = 0).

For the intercept, we want to detect a deviation of ± 0.35 mmol/L, which may be converted to:

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/SD_a = 0.35/0.06 = 5.8$$

According to Table 1, the sample size requirement is N <40. By squared extrapolation, we obtain the approximate value N = 19 [$40 \times (4/5.8)^2$]. Thus, in this example, the sample size requirement with regard to testing for intercept deviation from zero is less demanding than that of testing for a critical slope deviation.

PROPORTIONAL SD_as

Table 2 covers application of weighted Deming regression analysis in situations with proportional SD_as (constant CV_as) and includes intervals with range ratios extending from 2 to 100. The standardized slope deviation here is:

$$\Delta\beta_{st} = (\beta - 1)/CV_a$$

and the standardized intercept deviation from zero is:

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/(CV_a \cdot x_m)$$

CV_a should be expressed as a fraction, not a percentage, and x_m is the midpoint of the interval of interest (*Appendix*). Application of Table 2 presupposes that the analytical methods have identical CV_as and that the analyte values are uniformly distributed over the intervals of interest. The CV_a refers to single or duplicate sets of measurements as appropriate. Approximate sample

Table 2. Sample size table for comparison of methods with proportional SD_as using weighted Deming regression analysis.[a]

	Range ratios									
	2	**2.5**	**3**	**4**	**5**	**8**	**10**	**25**	**50**	**100**
$\Delta\beta_{st} = (\beta - 1)/CV_a$										
1	544	320	226	150	114	75	64	45	37	37
2	144	82	61	40	33	23	20	18	15	15
3	66	42	29	22	17	14	12	≤10	≤10	≤10
4	39	26	19	15	12	≤10	≤10	≤10	≤10	≤10
$\Delta\alpha_{0st} = (\alpha_0 - 0)/(CV_a \cdot x_m)$[b]										
1	521	281	180	99	69	34	26	≤10	≤10	≤10
2	130	70	45	28	20	14	11	≤10	≤10	≤10
3	60	32	24	15	12	≤10	≤10	≤10	≤10	≤10
4	35	22	15	≤10	≤10	≤10	≤10	≤10	≤10	≤10

[a] Necessary sample sizes for test of slope deviation from 1 or intercept deviation from zero by weighted Deming regression analysis given proportional SD_as, i.e., constant CVs: $CV_{ax} = CV_{ay} = CV_a$. Uniform distributions of x and y on intervals with the given range ratio. The range ratio is the maximum value divided by the minimum value of the considered interval. WLR (given CV_{ax} = 0) requires approximately one-half as many samples as the sample sizes listed. Type I error, 5%; power, 90%.

[b] x_m is the midpoint of the actual interval of interest, e.g., x_m equals 6 for the interval (4; 8), which has a range ratio of 2.

sizes for the application of WLR, assuming $CV_{ax} = 0$, are derived by division of the sample size values by two.

With regard to the testing of slope deviations, the strong impact on sample size of the range of values is apparent. The required number of samples is of the same order of magnitude as for analogous situations with constant SD_as. For example, detection of one standardized slope deviation with the given type I error and power requires >500 samples when the range ratio is 2, but only 37 samples for a range ratio of 50 or higher.

For deviations in the intercept, high sample size requirements are also present at low range ratios, e.g., 521 observations are necessary to detect one standardized deviation at a range ratio of 2, decreasing to ~70 at a range ratio of 5 and finally a negligible sample size at high range ratios. The required number of samples is the same order of magnitude as in analogous cases with constant SD_as. Notice that intercept deviations here are standardized with regard to the CV_a multiplied by x_m, which corresponds to the SD_a value in the constant case with the CV_a computed at the mean of the interval.

For standardized deviations that are not covered, the approximate sample size may be obtained by squared inter- or extrapolation. In the next section, the sample size requirement is considered for a comparison of two glucose methods with proportional SD_as.

planning a comparison of two glucose methods (proportional SD_as)

We first decide on the critical differences that should be detected. For convenience, we again apply CLIA 88 rules, which correspond to $\Delta = 10\%$ or 60 mg/L (6 mg/dL) at low levels. We consider a decision level corresponding to a fasting plasma glucose concentration of 1260 mg/L (126 mg/dL), which is a diagnostic limit for diabetes *(17)*. Again we notice that CLIA 88 rules relate to the total error. Assuming proportional SD_as with a CV_a of 3% for both methods, we obtain the critical systematic difference as:

$$\Delta_c = 10\% - 1.65 \cdot 3\% = 5\%$$

We have:

$$\Delta_c = Y_c - X_c = \alpha_0 + (\beta - 1) X_c$$

which corresponds to:

$$(0.05 \cdot 126) = Y_c - 126 = \alpha_0 + (\beta - 1) \cdot 126$$

or

$$6.3 = 132.3 - 126 = \alpha_0 + (\beta - 1) \cdot 126$$

This relationship translates to a requirement of detecting α_0 equal to ± 63 mg/L (± 6.3 mg/dL), if the systematic difference is related to an intercept deviation. Ascribing the systematic difference to a slope deviation implies a demand of detecting $\beta = 1.05$ or 0.95.

Let us now consider the following conditions: a measurement range of 600-3000 mg/L (60–300 mg/dL), i.e., a range ratio of 5; the midpoint of the interval (x_x) = 1800 mg/L (180 mg/dL); and for both methods, proportional SD_as with CV_as = 0.03, which means 0.02 for duplicate sets of measurements by each method ($0.03/\sqrt{2}$). For the regression procedure, we use the weighted Deming regression analysis with a significance level (type I error) of 5% and a statistical power of 90%.

We standardize the slope deviation:

$$\Delta\beta_{st} = (\beta - 1)/CV_a = 0.05/0.02 = 2.5$$

From the column corresponding to a range ratio of 5 in Table 2, we find the required sample size to be between 17 and 33 duplicate measurements for each method. Interpolation gives $N = 25$ [$17 \times (3/2.5)^2$]. If only single sets of measurements are performed, 50 samples are required. For WLR assuming $CV_{ax} = 0$, approximately one-half the number of samples are required ($N = 13$).

For the intercept, we want to detect a deviation of ± 63 mg/L (± 6.3 mg/dL), which may be converted to:

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/(CV_a \cdot x_m) = 6.3/(0.02 \cdot 180) = 1.7$$

From Table 2, we obtain $N = 24$ by squared interpolation, i.e., close to the number necessary for testing the critical slope deviation.

Other decision levels might also be considered, e.g., the nonfasting plasma glucose concentration limit of 2000 mg/L (200 mg/dL) as being diagnostic for diabetes *(17)*. In this proportional error model, the standardized slope deviation to be detected is the same at all decision levels, but the requirement for intercept detection varies with the level; therefore, the most demanding situations occur at low levels.

influence of target value distributions on sample size requirements

In the guidelines and examples outlined above, a basic assumption has been the uniform distribution of target values on the intervals characterized by a given range ratio. A uniform distribution of values throughout the range of interest is generally recommended in method comparison studies, but this ideal may not always be attained. If the comparison study is based on samples from a healthy population, the distributions of target values may be gaussian, e.g., for serum concentrations of electrolytes, or skewed. Furthermore, if samples are included from both healthy subjects and patients, skewed distributions usually arise. A common situation may be represented by a mixture of 75% of samples from healthy or nearly healthy subjects and 25% from diseased subjects, giving rise to a distribution skewed to the right. It is possible to outline some rough guidelines concerning these situations, which are common in real world examples.

For a gaussian distribution of target values, the major portion (95%) of the observations are located within ± 2 SD from the mean. Thus, the gaussian case may be compared to a situation with a uniform distribution spanning this interval. For the cases with constant SD_as, the necessary sample sizes read from Table 1 should be

multiplied by a factor of 1.3 to cover equivalent cases with gaussian target value distributions. This factor applies to both the slope and the intercept. For skewed distributions with 75% of the target values located on the lower half of the intervals, the factor for the slope is 1.4. No general correction factor can be applied for the intercept, but the necessary sample sizes extend from 1 to 1.3 times those listed for uniform target value distributions, with the highest factor adjustments pertaining to the narrowest intervals.

For cases involving proportional SD_as, the corrections factors are as follows. For gaussian distributions of target values, the sample size requirement as regard testing of the slope equals 1.3–1.5 times those listed in Table 2. For the intercept, 1–1.8 times the listed sample sizes are required. For the skewed target value distribution mentioned above, the correction factors range from 1.4 to 1.9 for the slope. For the intercept, the factors extend from 1.4 to 3. The highest correction factors for skewed distributions apply to the intervals with the highest range ratios.

Application of unweighted forms of regression analysis to cases involving proportional SD_as

According to current practices in method comparison studies, it is usual to apply unweighted forms of regression analysis, i.e., OLR and Deming analysis, although the SD_as vary with the measurement level, for example corresponding to a proportional relationship (constant CV_as). Thus, it is of interest to consider what happens in these situations. This subject has been addressed previously, but some supplementary aspects are considered here *(9)*.

Basically, OLR provides unbiased estimates of slope and intercept if the SD_a for x is zero, irrespective of whether the SD_a for y is constant or varies with the measurement level. In the same way, the Deming procedure provides unbiased estimates of slope and intercept when the SD_as vary, provided that their ratio is constant throughout the measurement range. This aspect is important and means that the estimates of slope and intercept generally are reliable in this frequently occurring situation. However, additional aspects are to be considered: the reliability of the associated statistical analysis, and the efficiency of the unweighted estimation procedures.

Two problems can occur when OLR is applied to real-life examples: the lack of consideration of measurement errors in x, and the variation of the SD_a. The first problem is well known and most significant at low range ratios, in which cases a biased slope estimate arises *(7, 9–10)*. Some authors have recommended that OLR may be applied when the correlation coefficient exceeds 0.975 or 0.99 *(18, 19)*. In these cases, however, the bias of the slope estimate has a consequence in that the type I error for the statistical analysis increases, i.e., the null hypothesis is rejected too frequently. This increase may amount to several fold and may cause the null hypothesis of a slope equal to one to be rejected far more frequently than anticipated from the nominal level of 5%.

The presence of a proportional SD_a for the y measurements also independently tends to increase the type I error for the test of slope deviation when the OLR procedure is used because the standard error of the slope is underestimated. The phenomenon is most pronounced for skewed target value distributions, in which cases the type I error may increase three- to fourfold, implying that the type I error becomes 15–20% compared with the nominal level of 5%. For uniform and gaussian target value distributions, the increase is up to 7.5% and 10%, respectively. Finally, the precision of slope and intercept estimations is lower than that provided by WLR in cases involving a proportional SD_a for y measurements. For a range ratio of 10, 2.3 times as many observations are required for estimation of the slope with a given precision compared with the WLR procedure. For the intercept, the factor is 3.9.

In unweighted Deming analysis, the associated statistical analysis is only slightly perturbed in cases involving proportional SD_as. For uniform and gaussian target value distributions, the type I error generally is unaffected. For the skewed target value distribution, the type I error may be slightly increased at sample sizes <100, but at higher sample sizes, the correct value of 5% is attained. These relationships refer to estimation of the standard error by the computerized jackknife principle as performed here (*Appendix*).

The major problem associated with application of the unweighted Deming analysis in cases involving proportional SD_as is the suboptimality of the unweighted approach (Table 3). For uniform distributions with range

Table 3. Comparison of sample sizes providing the same precision of slope and intercept estimates by unweighted and weighted Deming regression analysis in situations with proportional SD_as.

	Range ratios[a]						
	2	3	5	10	25	50	100
Sample size for weighted Deming regression analysis	100	100	100	100	100	100	100
Equivalent sample size for unweighted Deming regression analysis for slope testing	116	132	155	230	307	345	370
Equivalent sample size for unweighted Deming regression analysis for intercept testing	116	145	184	389	1067	1960	4650

[a] Uniform x and y distributions are supposed on intervals with the given range ratio.

ratios from 2 to 100, 1.2 to 3.7 times as many samples are needed to obtain the same precision of the slope estimate by the unweighted compared with the weighted approach. Thus, the larger the range ratio is, the more inefficient the unweighted method is. If one intends to apply the unweighted Deming procedure in a situation with proportional SD_as, the sample size values of Table 2 should be multiplied by the adjustment factors that may be derived from Table 3.

For the intercept, 1.2 to 3.9 times as many samples are required for the unweighted Deming procedure compared with the weighted Deming procedure for range ratios from 2 to 10. For higher range ratios, the efficiency of the unweighted procedure drops dramatically; therefore, as many as ~46-fold more samples are required to detect the same intercept deviation. Finally, it should be underlined that the relationships in Table 3 presuppose a true proportional SD_a relationship throughout the considered interval. At very large range ratios, such as may be observed for various hormones measured by immunoassay procedures, the SD_a value often tends to approach a constant plateau in the lowest part of the measurement range, which means that a true proportional relationship is not present over the entire range. In this case, the difference in efficiencies between unweighted and weighted procedures will be smaller than the difference shown in Table 3.

Discussion

The selected sample size in a method comparison study usually is based on conventions or protocols expressing general guidelines. From a practical point of view, guidelines are useful and necessary. For example, the NCCLS guideline EP-9A suggests measurement of 40 duplicate samples by each method when a new method is introduced in the laboratory as a substitute for an established one *(18)*. Additionally, it has also been proposed that a vendor of an analytical test system should have made a comparison study based on at least 100 samples measured in duplicate with each method. The principle of increased requirements for vendors appears reasonable. This initial validation should be comprehensive to disclose the performance of the assay system in detail. Then the requirement for the ordinary user may be more modest.

Although these general guidelines are useful, they may not be sufficient in the context of a detailed method evaluation, i.e., the vendor's or developer's evaluation. Additionally, assessment of reference methods within the hierarchy of definitive/reference methods may also pose special demands *(20)*. When measurements within a hierarchy that extends from a definitive method, through reference methods, to routine methods are compared, the amount of possible bias in each step should be defined as closely as possible. Here statistical power/sample size considerations become of relevance in the context of a rational method comparison design.

To assure that the necessary sample size is being estimated, basic information for the analytical methods, such as measurement range and SD_a, preferably throughout the measurement range, should be available, so that it can be decided whether constant or proportional SD_as are most likely in the given situation. This information usually is present because reproducibility frequently is evaluated very early for a new method, and the desirable measurement range is rather characteristic for a given analyte. Medically relevant method differences at selected levels should also be considered. As mentioned earlier, various sources may be consulted, and here the focus has been on differences decided by a regulative authority in the form of the CLIA 88 guidelines *(6)*. The next step is to convert the differences of interest to standardized intercept and/or slope deviations as described. Now Tables 1 or 2 may be consulted. Tables 1 and 2 do not cover all situations, but the use of inter- or extrapolation may extend the possibilities. Additionally, one may perform approximate factor adjustments taking into account nonuniform distributions of target values as mentioned earlier. Otherwise, specific calculations using the formulas in the *Appendix* may be performed, perhaps supplemented with simulations.

The focus has here been on how to detect slope or intercept deviations of given magnitudes and how to perform *t*-tests separately for slope and intercept deviations. As an alternative, a bivariate approach is possible with an outline of an elliptical joint confidence region for slope and intercept. In this context, it is possible to operate with sets of intercept and slope values that fulfill given critical deviations *(21)*.

In a method comparison study, the choice of an appropriate regression analysis procedure is important. It is advisable to consider whether constant or proportional SD_as or other relationships are present. Most frequently, two routine methods are compared. In this situation, analytical errors for both sets of measurement should be recognized. In case involving equal SD_as for x and y, the sample size requirement is doubled compared with the situation without analytical errors for x, which is not unexpected: the amount of random error or uncertainty is simply doubled. Application of Deming regression analysis presupposes that the ratio between the SD_as for the two methods is constant. This assumption is fulfilled if the SD_as are constant throughout the measurement range, as is assumed in the case involving constant analytical error. If the SD_as vary with the measurement level, the assumption may still be fulfilled provided that the same relationship with the measurement level applies for both methods. A common situation is the proportional SD_a case discussed here, corresponding to constant CV_as for the methods, which often occur with good approximation. Unless the measurement range is very narrow, the Deming regression procedure is generally robust toward nonconstant or misspecified SD_a ratios *(10, 22, 23)*. If duplicate sets of measurements for both methods are available, the SD_a ratio is estimated conveniently from the actual

data set. In cases involving proportional SD_as, the simple, unweighted form of the Deming procedure still provides an unbiased slope estimate, but as shown here and previously, the weighted Deming procedure is the most efficient approach *(9)*. For a range ratio of 10, the weighted approach requires less than one-half the number of samples to provide the same precision of the slope estimate as the unweighted Deming procedure. Furthermore, if a computerized resampling procedure such as the jackknife method is used for estimation of SD_as, the assumption of normality of analytical error distributions is not necessary *(12, 24)*.

In the planning phase, factors other than statistical power should be taken into account. The representativeness of samples is important. Samples from relevant patient categories should be included to help disclose possible interference phenomena. The calculated sample size requirement may then be regarded as a minimum demand from a statistical point of view, which may be modified to take other aspects into account. Furthermore, it is important that an internal quality-control system is in effect to assure that the methods to be compared are running in the in-control state. Comparisons of measurements preferably should be undertaken over several days, e.g., at least 5 days, to ensure that the method comparison does not become dependent on the performance of the methods in one particular analytical run *(18)*.

Sample size estimations seldom have been considered in the context of method comparison studies based on regression analysis. In a recent study, Hartmann et al. *(25)* used simulations to evaluate the power of OLR and the Deming procedure for selected data examples. Range ratios of 1.5, 5, and 15 were evaluated in relation to CV_as of 2% or 5%. The data examples are not exactly comparable with those of the present study, and the number of simulation runs for each parameter selection is somewhat small. However, for situations with constant SD_as (homoscedasticity), the authors found sample size estimates of the same order of magnitude as observed here. The authors considered only the unweighted forms of regression analysis.

Passing and Bablok *(26)* studied the sample size required to obtain a given power by nonparametric rank regression analysis. For a series of simulation examples, these authors considered the necessary sample sizes for detection of given slope deviations with a power of 80% given a type I error of 5%, assuming proportional SD_as. The same pattern as observed here for testing of slope deviations was apparent: the larger the range of values, the smaller the required sample size (other factors being equal). Although the examples used by Passing and Bablok *(26)* are not exactly comparable to the situations dealt with here, it is apparent that the sample size requirements of the rank procedure exceed those of the weighted Deming procedure, which agrees with the fact that the latter procedure is more efficient than the rank procedure *(9)*.

In recent years, the difference plot described by Bland and Altman *(27)* has gained increasing popularity as a tool for evaluation of method comparison data *(28)*. Although the difference plot in itself is very instructive for displaying differences, the associated summary statistic in the form of a paired *t*-test is not appropriate for the analysis of method comparison data because it might be misleading in the presence of a systematic proportional difference *(29–31)*. The *t*-test only evaluates whether the mean measurement levels of two methods agree, but not whether the measurements are comparable throughout the measurement range. For example, if measurements in the low range by a new method tend to be higher than those of the established method, and vice versa in the high range, the averages of the measurements by each method agree, and the paired *t*-test shows no difference. A regression analysis, on the other hand, would clearly disclose the different performances of the methods.

In conclusion, the planning of a method comparison study to achieve a given power for detection of medically significant differences should be considered carefully. In this way, a method comparison study is likely to be conclusive: Either the null hypothesis of no difference is accepted, or the presence of a relevant difference is established. Otherwise, a statistically nonsignificant slope deviation from unity and/or intercept deviation from zero may either imply that the null hypothesis is true or be an example of a type II error, i.e., an overlooked real difference of medical importance.

References

1. Skendzel LP, Barnett RN, Platt R. Medically useful criteria for analytic performance of laboratory tests. Am J Clin Pathol 1985; 83:200–5.
2. Fraser CG, Hyltoft Petersen P. Desirable standards for laboratory tests if they are to fulfill medical needs. Clin Chem 1993;39: 1447–55.
3. Linnet K. Choosing quality control systems to detect maximum medically allowable analytical errors. Clin Chem 1989;35:284–8.
4. Laboratory Standardization Panel of the National Cholesterol Education Program. Current status of blood cholesterol measurements in clinical laboratories in the United States: a report from the Laboratory Standardization Panel of the National Cholesterol Education Program. Clin Chem 1988;34:193–201.
5. Stein EA, Myers GL. National Cholesterol Education Program recommendations for triglyceride measurement: executive summary. Clin Chem 1995;41:1421–6.
6. US Department of Health and Human Services. Medicare, Medicaid, and CLIA programs: regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Final rule. Fed Regist 1992;57:7002–186.
7. Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem 1979;25: 432–8.
8. Parvin CA. A direct comparison of two slope-estimation techniques used in method-comparison studies. Clin Chem 1984;30:751–4.
9. Linnet K. Evaluation of regression procedures for methods comparison studies. Clin Chem 1993;39:424–32.

10. Linnet K. Performance of Deming regression analysis in case of a misspecified analytical error ratio. Clin Chem 1998;44:1024–31.
11. Ross JW, Lawson NS. Analytical goals, concentration relationships, and the state of the art for clinical laboratory precision. Arch Pathol Lab Med 1995;119:495–513.
12. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med 1990;9:1463–73.
13. Hald A. Statistical theory with engineering applications. New York: Wiley, 1952:551–7.
14. Snedecor GW, Cochran WG. Statistical methods, 6th ed. Ames, IA: Iowa State University Press, 1967:139–67.
15. Linnet K. CBstat: a program for statistical analysis in clinical biochemistry. Reference manual. Risskov, Denmark: K Linnet, 1998:1–53.
16. Snedecor GW, Cochran WG. Statistical methods, 6th ed. Ames, IA: Iowa State University Press, 1967:113 pp.
17. American Diabetes Association. Screening for type 2 diabetes. Diabetes Care 1998;21(Suppl 1):S20–2.
18. National Committee for Clinical Laboratory Standards. Method comparison and bias estimation using patient samples. Approved guideline. NCCLS document EP9-A. Villanova, PA: NCCLS, 1995; 15(17):1–36.
19. Wakkers PJM, Hellendoorn HBA, Op De Weegh GJ, Heerspink W. Applications of statistics in clinical chemistry. A critical evaluation of regression lines. Clin Chim Acta 1975;64:173–84.
20. Tietz NW. A model for a comprehensive measurement system in clinical chemistry. Clin Chem 1979;25:833–9.
21. Gerbet D, Auget J-L, Maccario J, Cazalet C, Raichvarg D, Ekindjian OG, Yonger J. New statistical approach in biochemical method-comparison studies by using Westlake's procedure, and its application to continuous-flow, centrifugal analysis, and multilayer film analysis techniques. Clin Chem 1983;29:1131–6.
22. Beech DG. Some notes on the precision of the gradient of an estimated straight line. Appl Stat 1961;10:14–31.
23. Lakshminarayanan MY, Gunst RF. Estimation of parameters in linear structural relationships: sensitivity to the choice of the ratio of error variances. Biometrika 1984;71:569–73.
24. Wu CFJ. Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). Ann Stat 1986;14:1261–95.
25. Hartmann C, Smeyers-Verbeke J, Penninckx W, Massart DL. Detection of bias in method comparison by regression analysis. Anal Chim Acta 1997;338:19–40.
26. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. J Clin Chem Clin Biochem 1984; 22:431–45.
27. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
28. Petersen PH, Stöckl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, et al. Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots. Clin Chem 1997;43:2039–46.
29. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method comparison studies. Clin Chem 1973;19: 49–57.
30. Westgard JO, deVos D, Hunt MR, Quam EF, Carey RN, Garber CC. Concepts and practices in the evaluation of clinical chemistry methods. Part III. Statistics. Am J Med Technol 1978;44:552–70.
31. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–68.
32. Kendall MG, Stuart A. The advanced theory of statistics, Vol. 2. London: Charles Griffin, 1973:393–7.

Appendix

SAMPLE SIZE FORMULA

A general, simplified formula for computation of the approximately necessary sample size for detection of adifference Δ with a given type I error and power *(16)* is:

$$N = (c/\Delta)^2(t_{p/2} + t_{1-q})^2 \quad (1)$$

where c is the proportionality factor determining the standard error of the estimated difference D, which corresponds to the true difference Δ, i.e., the standard error of D is $c/\sqrt{N}$. $t_{p/2}$ is the t value for the given type I error (significance level). For the usual level of $P = 5\%$, $t_{p/2} = 1.96$ (asymptotically). t_{1-q} is the t value for the desired power $1 - q$, where q is the type II error. For a power of 90%, t_{1-q} is 1.28 (asymptotically).

REGRESSION ANALYSIS GIVEN CONSTANT SD_aS

This method applies to OLR and Deming regression analysis. For the OLR procedure, the slope, intercept, and their standard errors are estimated *(14)* as:

$$x_m = \sum x_i/N;\ y_m = \sum y_i/N$$

$$u = \sum(x_i - x_m)^2;\ q = \sum(y_i - y_m)^2;$$

$$p = \sum(x_i - x_m)(y_i - y_m)$$

$$b = p/u;\ a_0 = y_m - bx_m$$

$$Y_{esti} = a_0 + bx_i$$

$$SD_{y.x} = [\sum(y_i - Y_{esti})^2/(N - 2)]^{0.5}$$

$$SE(b) = SD_{y.x}/\sqrt{u};\ SE(a_0) = SD_{y.x}[(1/N) + (x_m^2/u)]^{0.5}$$

Y_{esti} refers to the estimated Y value for a given x_i according to the regression equation. In cases involving duplicate sets of measurements, each x_i and y_i represent the mean of individual measurements [$x_i = (x_{1i} + x_{2i})/2$ and $y_i = (y_{1i} + y_{2i})/2$].

Given a uniform distribution of x, for any given interval $u_{st} = u/N$ is nearly constant over a wide range of N values. Furthermore, the expression x_m^2/u_{st} is scale invariant; it depends only on the ratio between the maximum and minimum values of the interval, i.e., the range ratio. On the basis of this background, we rewrite the standard error expressions as:

$$SE(b) = SD_{y.x}/\sqrt{u} = CV_{ay}/(u/x_m^2)^{0.5}$$

$$= (1/\sqrt{N})\ CV_{ay}\ (x_m^2/u_{st})^{0.5}$$

$$= (1/\sqrt{N})\ CV_{ay}\ c_b$$

$$SE(a_0) = SD_{y.x}(1/N + x_m^2/u)^{0.5}$$

$$= (1/\sqrt{N})\ SD_{ay}\ (1 + x_m^2/u_{st})^{0.5}$$

$$= (1/\sqrt{N})\ SD_{ay}\ c_{a0}$$

where c_{a0} and c_b are constants depending only on the range ratio. Under the given model, $SD_{y.x}$ equals the

analytical SD of method y (SD_{ay}). CV_{ay} here is the CV at the middle of the interval in question, i.e., SD_{ay}/y_m. We here suppose that there is only a negligible difference between x_m and y_m; therefore, we have that CV_{ay} is approximately equal to SD_{ay}/x_m. Thus, if the slope deviation from unity and the intercept deviation from zero are expressed in standardized forms, i.e., in CV_a and SD_a units of method y, respectively, we obtain the following sample size formulas analogous to Eq. 1:

$$N_{slope} = (c_b/\Delta\beta_{st})^2(t_{p/2} + t_{1-q})^2 \text{ and} \quad (2)$$

$$\Delta\beta_{st} = (\beta - 1)/CV_a$$

$$N_{intercept} = (c_{a0}/\Delta\alpha_{0st})^2(t_{p/2} + t_{1-q})^2 \text{ and} \quad (3)$$

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/SD_a$$

The Deming regression line is estimated as:

$$b = \{(\lambda q - u) + [(u - \lambda q)^2 + 4\lambda p^2]^{0.5}\}/2\lambda p;$$

$$\lambda = SD_{ax}^2/SD_{ay}^2$$

$$a_0 = y_m - bx_m$$

$$Y_{esti} = a_0 + bX_{esti}$$

SD_as are estimated from duplicate sets of measurements as:

$$SD_{ax}^2 = (1/2N)\sum(x_{2i} - x_{1i})^2 \text{ and}$$

$$SD_{ay}^2 = (1/2N)\sum(y_{2i} - y_{1i})^2$$

For the Deming procedure, general formulas for standard errors of slope and intercept are complicated *(32)*, and in practice, they are most easily estimated using a computerized resampling principle such as the jackknife method *(12, 15)*. However, for a standard situation with a slope close to unity and equal SD_as for methods x and y, i.e., $SD_{ax} = SD_{ay} = SD_a$ and $CV_a = SD_a/x_m$, we have the following relationships:

$$SE(b) = \sqrt{2}\,(1/\sqrt{N})\,CV_a\,c_b$$

$$SE(a_0) = \sqrt{2}\,(1/\sqrt{N})\,SD_a\,c_{a0}$$

This relationship has been verified by simulations. Thus, under the present standardized conditions, Eqs. 2 and 3, developed for the OLR method, also apply to the Deming procedure with the modification that the sample size values are multiplied by two. On this basis, Table 1 has been constructed such that the stated sample size values apply to Deming regression analysis. For the OLR procedure, the sample sizes are divided by two. At parameter combinations corresponding to small to moderate sample sizes (<100–150), the computations were supplemented with simulations (1000 runs for each parameter combination), and sample sizes were adjusted. Accordingly, there is not an exact inverse relationship between the standardized Δ value and the squared sample size.

REGRESSION ANALYSIS GIVEN PROPORTIONAL SD_aS

This situation applies to WLR and weighted Deming regression analysis.

For WLR *(9, 13)*, we have:

$$x_{mw} = \sum w_i x_i/\sum w_i;\ y_{mw} = \sum w_i y_i/\sum w_i$$

$$u_w = \sum w_i(x_i - x_{mw})^2;\ q_w = \sum w_i(y_i - y_{mw})^2;$$

$$p_w = \sum w_i(x_i - x_{mw})(y_i - y_{mw})$$

$$b = p_w/u_w;\ a_0 = y_{mw} - bx_{mw}$$

$$Y_{esti} = a_0 + bx_i$$

The weights (w_i) are inversely proportional to the squared SD_a of y measurements at a given level. The SD_a is assumed to be a function $[h(\cdot)]$ of x:

$$SD_{ay} = kh(x_i);\ w_i = [h(x_i)]^{-2}$$

where k is a proportionality factor. For the proportional SD_a case, which is assumed to hold true in the present context, $w_i = 1/x_i^2$. The proportionality factor is estimated from the dispersion around the line:

$$k = [\sum(y_i - Y_{esti})^2 w_i/(N - 2)]^{0.5}$$

The standard errors of slope and intercept are:

$$SE(b) = k/\sqrt{u_w};\ SE(a_0) = k(1/\sum w_i + x_{mw}^2/u_w)^{0.5}$$

For situations with an intercept close to zero and a slope close to one, the factor k is approximately equal to the CV_a for y measurements CV_{ay}. Given a uniform x distribution, it turns out that $u_{wst} = u_w/N$ is scale invariant and is nearly constant for a wide range of N. Thus, we have approximately:

$$SE(b) = k/\sqrt{u_w} = CV_{ay}/\sqrt{u_w} = (1/\sqrt{N})\,CV_{ay}/\sqrt{u_{wst}} = (1/\sqrt{N})\,CV_{ay}\,c_b$$

$$SE(a_0) = k[(1/\sum w_i) + (x_{mw}^2/u_w)]^{0.5} = (1/\sqrt{N})\,CV_{ay}[(1/\sum_{st} w_i) + (x_{mw}^2/u_{wst})]^{0.5} = (1/\sqrt{N})\,CV_{ay}\,c_{a0}$$

where $c_b = 1/\sqrt{u_{wst}}$ and $c_{a0} = x_{mw}\,[1/(x_{mw}^2\Sigma_{st}w_i) + 1/u_{wst}]^{0.5}$, with $\Sigma_{st}w_i = (1/N)\Sigma w_i$. c_b is scale invariant. For c_{a0}, the expression in brackets is scale invariant, whereas x_{mw} is proportional to the scale for a given range ratio in the same way as the ordinary mean x_m. This proportionality is taken into account in the expression for $\Delta\alpha_{0st}$ as displayed below, where for convenience x_m enters the

formula instead of x_{mw} and c_{a0} is adjusted accordingly. The sample size formulas are:

$$N_{slope} = (c_b/\Delta\beta_{st})^2(t_{p/2} + t_{1-q})^2 \text{ and} \quad (4)$$

$$\Delta\beta_{st} = (\beta - 1)/CV_a$$

$$N_{intercept} = (c_{a0}/\Delta\alpha_{0st})^2(t_{p/2} + t_{1-q})^2 \text{ and} \quad (5)$$

$$\Delta\alpha_{0st} = (\alpha_0 - 0)/(CV_a \cdot x_m)$$

For the weighted Deming procedure, the slope and intercept are estimated as:

$$b = \{(\lambda q_w - u_w) + [(u_w - \lambda q_w)^2 + 4\lambda p_w{}^2]^{0.5}\}/2\lambda p_w$$

$$a_0 = y_{mw} - bx_{mw}$$

$$Y_{esti} = a_0 + bX_{esti}$$

It is supposed here that the ratio between the squared SD_as is constant ($\lambda = SD_{ax}{}^2/SD_{ay}{}^2$) throughout the measurement range. The SD_as are functions of the target values:

$$SD_{ax} = k_x h_x(X_i) \text{ and } SD_{ay} = k_y h_y(Y_i)$$

Under the assumption of proportional SD_as and a slope close to one and an intercept close to zero, λ is approximately equal to $CV_{ax}{}^2/CV_{ay}{}^2$. In weighted Deming regression analysis, the weights w_i, which equal $1/[(X_{esti} + \lambda Y_{esti})/(1 + \lambda)]^2$, are obtained by an iterative principle as described using the CBstat program *(12, 15)*. The weights are expressed here as a weighted mean of X_{est} and Y_{est}, which is slightly more optimal than the simple mean described earlier *(9, 12)*.

For the weighted Deming procedure, general formulas for the standard errors of the slope and intercept are complicated, and again the jackknife method may be applied *(12, 15)*. As above, however, for the simplified situation considered here with a slope close to unity and equal CV_as for methods x and y, i.e., $CV_{ax} = CV_{ay} = CV_a$, we have the following relationships:

$$SE(b) = \sqrt{2}(1/\sqrt{N})CV_a c_b$$

$$SE(a_0) = \sqrt{2}(1/\sqrt{N})CV_a c_{a0}$$

Table 2 has been constructed primarily from these formulas, with adjustments based on simulation studies on the weighted Deming procedure. The sample sizes for WLR are approximately one-half of the sample sizes listed. The weighted approach presupposes positive sample values.