JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 14, No. 4, pp. 881–892, 2004

Bias Estimation in Method Comparison Studies

Robert T. Magari*

Beckman Coulter Corporation, Miami, Florida, USA

ABSTRACT

A test and a reference analytical method are usually compared for agreement based on paired data obtained from several independent subjects. Bias between two methods can be classified as constant and proportional. In this article, we provide an approach for maximum likelihood estimation of total bias between two methods and partitioning it into constant and proportional bias for each subject. Normal, binomial, or Poisson distribution are the conditional distributions of the response variable that we have considered here, whereas subjects are considered to be random sample from a normally distributed population. Real data on blood cell counts and hemoglobin are used for demonstration. The estimate of biases can be used to test different statistical hypotheses and/or for graphical interpretation of the agreement. The partitioning of total biases in terms of constant and proportional gives an insight on the sources of disagreement between two methods and helps designers and manufacturer's define a remedial strategy.

Key Words: Bias; Maximum likelihood estimation; Normal distribution; Binomial distribution; Poisson distribution.

*Correspondence: Robert T. Magari, Beckman Coulter Corporation, 11800 SW 147 Ave., Miami, FL 33116-9015, USA; Fax: 305-380-3699; E-mail: robert_magari@coulter.com.

DOI: 10.1081/BIP-200035450 1054-3406 (Print); 1520-5711 (Online)

INTRODUCTION

It is usually required that a newly released clinical laboratory and/or diagnostic method be compared with the one that is currently in use with regard to precision and accuracy. Although the current method may not be the best, it is generally called the reference, and the new method being tested is called the test. This comparison is usually based on paired data obtained from several independent subjects, sometimes measured in more than one replicate. The test method agrees with the reference if the bias between them is not relevant, whereas both of them are considered to measure with a certain degree of imprecision. Disagreement bias between two methods may come from different sources but in general is classified as constant and proportional (Westgard and Hunt, 1972). Constant bias (*CB*) is a form of systematic deviations estimated as an average difference between the test and the reference method. Two methods may agree on the average, but they may exhibit biases in particular range(s) of measurements. These biases are referred to as proportional (*PB*). The *PB* is regarded as a significant difference between two methods for each subject that is independent of *CB*. Several other authors have proposed more detailed breakdowns of *PB* that can be related to specific cases and protocols (Krouwer, 1972).

Several statistical approaches are used to study agreement between two analytical methods. However, no single statistical parameter can correctly describe all the aspects of agreement between a test and a reference method and adequately estimate biases of different sources. Linear regression is probably the most popular technique used in practice, although when slope is statistically different from one, none of the linear regression parameters can be related to *CB* and *PB* (Magari, 2002). Ordinary least square (OLS) estimates of slope and intercept are easily obtained even with common nonstatistical software. Other authors prefer orthogonal least squares estimation (Deming regression) because both methods are measured with error, whereas Passing and Bablok have provided an distribution free nonparametric approach as well (Cornbleet and Gochman, 1979; Linnet, 1993; and Passing and Bablok, 1983). Correlation coefficient-type approaches based on a bivariate normal distribution of the data are also given in the literature (Lin, 1989; St. Laurent, 1998; Bartko, 1994; Hutson et al., 1998). An omnibus test for simultaneous comparison of accuracy and precision is proposed in Blackwood and Bradley (1991).

Most of the statistical models used in method comparison studies are designed for normally distributed data. However, some systems in clinical diagnostic are based on counting of certain particles rather than measuring a substance. In some of these cases, particularly in hematology where counting of cell types is of primary importance, the assumption of normal distribution is not always appropriate. Thus, other distributions beside normal distribution need to be considered. In this article we present a mixed-effect model for maximum likelihood estimation of total biases between two methods and partitioning it into *CB* and *PB* for each subject. Furthermore, we extend the assumption of conditional normal distribution of the response variable to binomial and Poisson distributions, which may be more appropriate for certain types of count data. Finally, we present several application examples using real data.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

ESTIMATION

Model for One Analytical Method

Let ϕ_i be the expected outcome for each subject measured by a certain analytical method ($1 \leq i \leq s$). The observed jth replicate of this outcome ($1 \leq j \leq n$) can be expressed as

$$Y_{ij} = \phi_i + \varepsilon_{ij} \tag{1}$$

where ε_{ij} is a random error term related to the imprecision of the measurements. The number of replicates (n) may vary from subject to subject and from analytical method to analytical method. There are cases when $n = 1$. Let ϕ_i be further partitioned as

$$\phi_i = \mu + u_i \tag{2}$$

with μ being an overall effect of the method and u_i a random effect associated with each subject. We assume subjects to be a random sample from a normally distributed population of interest, and $u_i \sim \text{NOR}(0, \sigma_u^2)$. Thus, the probability density function of u_i can be written as

$$q(u_i \mid \sigma_u^2) = (2\pi\sigma_u^2)^{-1/2}\exp\left(\frac{u_i^2}{2\sigma_u^2}\right). \tag{3}$$

Let express the joint probability density function of Y_{ij} and u_i as,

$$p(Y_{ij} \mid \boldsymbol{\varphi}, u_i)q(u_i \mid \sigma_u^2), \tag{4}$$

where $p(Y_{ij} \mid \boldsymbol{\varphi}, u_i)$ represents the conditional distribution of Y_{ij}, and $\boldsymbol{\varphi}$ is the vector of the unknown parameters.

The conditional distribution of Y_{ij} depends on the types of data collected and is related to the nature of the outcome variable, technology, measuring systems, etc. In this article we are considering the following three distributions,

- Normal distribution

  $Y_{ij} \sim \text{NOR}(\phi_i, \sigma_\varepsilon^2)$, $\boldsymbol{\varphi} = (\phi_i, \sigma_\varepsilon^2)$ and

  $$p(Y_{ij} \mid \phi_i, \sigma_\varepsilon^2, u_i) = (2\pi\sigma_\varepsilon^2)^{-1/2}\exp\left(\frac{(Y_{ij} - \phi_i)^2}{2\sigma_\varepsilon^2}\right) \tag{5}$$

- Binomial distribution

  $Y_{ij} \sim \text{BIN}(n_{ij}, \phi_i), \boldsymbol{\varphi} = (n_{ij}, \phi_i)$ and

  $$p(Y_{ij} \mid n_{ij}, \phi_i, u_i) = \left(\frac{n_{ij}}{Y_{ij}}\right)\phi_i^{Y_{ij}}(1-\phi_i)^{n_i - Y_{ij}} \tag{6}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

- Poisson distribution

$$Y_{ij} \sim \text{POI}(\phi_i), \boldsymbol{\varphi} = \phi_i \text{ and}$$

$$p(Y_{ij} \mid \phi_i, u_i) = \frac{\exp(-\phi_i)\phi_i^{Y_{ij}}}{Y_{ij}!} \tag{7}$$

The estimation of the parameters of Eq. (1) is needed for assessing biases and agreement between methods as will be elaborated in the following section. The estimation of $\boldsymbol{\varphi}$ and σ_u^2 can be achieved by numerically maximizing the marginal likelihood function

$$L(\boldsymbol{\varphi}, \sigma_u^2) = \prod_{i=1}^{s} \int p(Y_{ij} \mid \boldsymbol{\varphi}, u_i)\ q(u_i \mid \sigma_u^2) \mathrm{d}u_i \tag{8}$$

Because the above integral does not have a closed form, several authors have proposed approximations that perform efficiently under certain conditions. We used the adaptive Gaussian quadrature method that generally performs well for the type of data, models, and the distributions considered in this paper (Pinheiro and Bates, 1995). The main steps of this approach are described below. For more details and motivation the reader is referred to Pinheiro and Bates (1995), SAS/Stat User's Guide (2000), and Lindstrom and Bates (1998).

First, predefined values of Gauss–Hermite abscissa and weights (z_k, w_k; $1, \ldots, n$) are obtained for the most common kernels as proposed by Golub (1973) for each kth quadrature point. Then, these values besides starting values for the parameters are used to iteratively maximized the following approximation

$$\approx \sum_{k=1}^{n} \left[p(Y_{ij} \mid \boldsymbol{\varphi}, \hat{\boldsymbol{u}}_i + \Gamma(\boldsymbol{\varphi}, \sigma_u^2)^{-1/2} z_k)\ q(\hat{\boldsymbol{u}}_i + \Gamma(\boldsymbol{\varphi}, \sigma_u^2)^{-1/2} z_k \mid \sigma_u^2) W_k \right] \tag{9}$$

In the above equation, $\hat{\boldsymbol{u}}_i$ are the empirical Bayes estimates of u_i, and $\Gamma(\boldsymbol{\varphi}, \sigma_u^2)$ is the Hessian matrix from the empirical Bayes optimization. Furthermore, Newton–Raphson algorithm is used for optimization (Lindstrom and Bates, 1998). Standard errors of the estimates of $\boldsymbol{\varphi}$ are obtained from the diagonal of the inverse of the Hessian matrix, $\widehat{\mathbf{H}}(\boldsymbol{\varphi})$, whereas the standard error of the estimates of u_i are obtained from the diagonal of the matrix calculated as follows

$$\Gamma^{-1}(\boldsymbol{\varphi}, \sigma_u^2)^1 + \left(\frac{\partial \hat{\boldsymbol{u}}_i}{\partial \boldsymbol{\varphi}}\right) \widehat{\mathbf{H}}^{-1}(\boldsymbol{\varphi}) \left(\frac{\partial \hat{\boldsymbol{u}}_j}{\partial \boldsymbol{\varphi}}\right)^T \tag{10}$$

Model for Biases Between Two Analytical Methods

Let a number of s subjects be randomly analyzed by the two methods. Let ϕ_{Ti} and ϕ_{Ri} be the expected outcomes for each subject for the test and the reference methods, respectively. Then the expected value of the difference between two

methods for each subject can be expressed as

$$D_i = \phi_{Ti} - \phi_{Ri} = (\mu_T + u_{Ti}) - (\mu_R + u_{Ri}) \tag{11}$$

where the subscripts T and R of μ and u_i, defined similarly as in Eq. (1), stand for the test and the reference methods. By rearranging the terms of the above equation we obtain

$$D_i = (\mu_T - \mu_R) + (u_{Ti} - u_{Ri}) \tag{12}$$

The first term is the average difference between two methods that represents constant bias (CB). The second term, which is a vector of s elements ($s =$ number of subjects), represents the proportional bias (PB_i). D_i is generally referred to as total bias. Thus,

$$D_i = CB + PB_i \tag{13}$$

We consider proportional biases of different subjects measured by the same method to be independent, $\text{cov}(PB_i,\ PB_i') = 0$, because each subject is randomly analyzed by each method and any types of carryover effects are insignificant. The observed differences between two methods for each subject will also have an error term, which is related to imprecision of measurement and/or other sources of random errors (IP_i) and depend on the conditional distribution of the outcome variable (Houwen, 1990). Thus, the observed difference can be expressed as

$$\overline{Y}_{Ti} - \overline{Y}_{Ri} = D_i + IP_i = CB + PB_i + IP_i \tag{14}$$

where, $\overline{Y}_{Ti}$ and $\overline{Y}_{Ri}$ are average observed values for each subject measured by the test and reference method, respectively. The estimates of CB and PB_i can be obtained as

$$\widehat{CB} = \hat{\mu}_T - \hat{\mu}_R \tag{15}$$

$$\widehat{PB}_i = \hat{u}_{Ti} - \hat{u}_{Ri} \tag{16}$$

Standard errors of these estimates are

$$se(\widehat{CB}) = \sqrt{se(\hat{\mu}_T)^2 + se(\hat{\mu}_R)^2 - 2r_{T,R}se(\hat{\mu}_T)se(\hat{\mu}_R)} \tag{17}$$

$$se(\widehat{PB}_i) = \sqrt{\frac{se(\hat{u}_{Ti})^2 + se(\hat{u}_{Ri})^2 - 2r_{T,R}se(\hat{u}_{Ti})se(\hat{u}_{Ri})}{n}} \tag{18}$$

$r_{T,R}$ is the correlation coefficient between $\overline{Y}_{Ti}$ and $\overline{Y}_{Ri}$, and n is the number of replicates. The values of $\hat{\mu}_T, \hat{\mu}_R, \hat{u}_{Ti}, \hat{u}_{Ri}, se(\hat{\mu}_T), se(\hat{\mu}_R), se(\hat{u}_{Ti})$, and $se(\hat{u}_{Ri})$ are obtained from the maximum likelihood estimation process of the model in Eq. (1)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

for each analytical method separately. The estimate of total bias and its standard error are obtained as

$$\widehat{D}_i = \widehat{CB} + \widehat{PB}_i \tag{19}$$

$$se(\widehat{D}_i) = \sqrt{se(\widehat{CB})^2 + se(\widehat{PB}_i)^2} \tag{20}$$

Standard errors are used to calculate the approximate $(1-\alpha)100$ confidence intervals for D_i, where α is the level of significance.

DATA COLLECTION AND ANALYSIS

Data used in this article for demonstrations are collected from four different experiments (sets 1–4 in Table 1) conducted at Beckman Coulter Inc. Hematology Laboratory in Miami, Florida. Datasets consist of several randomly selected and independent blood specimens analyzed by a reference and a test method in replicates during the same setup. Reference and test methods that represented different instruments and/or instrument settings were not the same for all sets. The response variables for each set are described below.

Set 1—leukocytes (WBCs). The WBCs are types of blood cells that protect the body from disease agents and other foreign substances in the bloodstream. Instruments count the number of WBCs in a suspension of blood cells belonging to a subject (specimen) and express this number per one micro liter (μL) of volume. Counts for each subject are considered to be POI(ϕ_i).

Set 2—monocytes (MO) and Set 3—neutrophils (NE). These types of WBCs are involved in the body's defense systems and are responsible for several phagocytic functions. Instruments count the number of MO and NE for a certain total number of WBCs and express their ratio in percentage. The number of MO and NE for each subject are considered to be BIN(n_{ij}, ϕ_i), where n_{ij} is the WBC count of the jth replicate of the ith subject and ϕ_i is the proportion of MO and NE for that subject.

Table 1. Description of datasets and estimation of constant biases for each of them.

Set	Measured variable	Units	Conditional distribution	No. of subjects	Constant bias with 95% confidence interval		
					Estimate	Lower	Upper
1	Leukocytes	cells/μL	Poisson	35	142	83	201
2	Monocytes	%	Binomial	60	−0.46	−0.66	−0.27
3	Neutrophils	%	Binomial	60	0.07	−0.26	0.39
4	Hemoglobin	g/dL	Normal	20	0.72	0.42	1.02

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Set 4—hemoglobin (HGB). The HGB is an iron-protein compound in red blood cells that gives blood its red color and transports oxygen, carbon dioxide, and nitric oxide. The amount of HGB measured for each specimen and expressed in grams per deci liter (g/dL) is considered to be NOR(ϕ_i, σ_ε^2).

Data were electronically collected from the instruments and converted to an SAS® 8.01 (SAS Institute Inc., Cary, NC) format for statistical analysis. PROC MLMIXED of SAS is used to obtain the estimates of Eq. (1). The response variable is defined in terms of the respective distributional parameter in the MODEL statement of PROC MLMIXED, whereas u_i are defined in the RANDOM statement as NOR$(0, \sigma_u^2)$ clustered according to subjects. Program default values for the optimization, convergence, and termination criteria were not changed (SAS/Stat User's Guide, 2000). In the Set 1 data (Poisson distribution), the algorithm converged only after we performed a grid search of initial values for the model parameters. Algorithm converged much easier for the other distributions. Readers who are interested in the SAS software may contact the author.

RESULTS

The estimates of *CB* along with their 95% confidence intervals are given in Table 1. *CB* in sets 1, 2 and 4 are clearly different from zero. However, the magnitude of *CB* for MO (set 2) does not affect any diagnostic decision because a difference of smaller than 1% is usually considered as clinically nonrelevant in healthy individuals (Jones et al., 1996). The *CB* for NE is statistically as well as clinically nonsignificant. Meaningful differences are obtained for HGB and WBCs. The test method in set 4 measures an average HGB of 0.72 g/dL higher than the reference. This difference is statistically as well as clinically relevant and substantiates the fact that the two methods do not agree with each other.

The estimates of total biases for some selected subjects are shown in Table 2. To make comparisons from a measurement range perspective, the numbers within each set are sorted on the basis of the value of the averages between two methods. Typical subjects that represent the overall trend(s) are shown. Estimates of D_i are calculated on the basis of Eq. (13) for each subject. For instance, Subject 1 in Set 1, $\widehat{CB} = 142$ cells/μL, $\widehat{PB}_1 = -8$ cells/μL, and $\widehat{D}_1 = \widehat{CB} + \widehat{PB}_1 = 142 + (-8) = 134$ cells/μL. The difference of 1 cells/μL between $\widehat{D}_1$ and the observed difference between two methods for that subject (135 cells/μL) can be attributed to the imprecision of measuring device as well as other sources of random errors. Standard error of the estimated difference for this subject is, $se(\widehat{D}_1) = 5.1$ cells/μL. Considering the distribution of $\widehat{D}_i$ to be approximately normal, the ratio $z_c = \widehat{D}_i / se(\widehat{D}_i)$ provides a test for $H_0 : D_i = 0$. For Subject 1, Set 1, $z_c = \widehat{D}_1 / se(\widehat{D}_1) = 26.27$ and p-value, $2P(z \geq 26.27) \approx 0$. The-values for $H_0 : D_i = 0$ in the selected subjects are shown in the last column of Table 2. Other hypotheses can be tested as well. Test statistics for comparison of biases in different ranges are also possible because $\widehat{D}_i = 0$ are independent, and the differences $\widehat{D}_i - \widehat{D}_{i'}$ can be considered as approximately normal variables with $se(\widehat{D}_i - \widehat{D}_i) = \sqrt{se(\widehat{D}_i)^2 + se(D_{i'})^2}$. Using Table 2 data, Set 2, the calculated test statistics for $H_0 : D_{i'} = D_4$ is $z_c = 9.9$ and $2P(z \geq 9.9) \approx 0$, whereas

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 2. Estimates of biases in some selected subjects.

Set	Constant bias	Subject ID	Average between methods	Observed difference	Proportional bias	Total bias		
						Estimate	Standard error	p-value for H_0: $D_i = 0$
1	142	1	5528	135	−8	134	29.9	0.0000
		25	6268	185	35	177	30.0	0.0000
		11	6568	−45	−144	−2	30.0	0.7390
		31	6708	305	129	271	30.0	0.0000
2	−0.0046	1	0.0847	−0.0006	0.0019	−0.0028	0.0014	0.0491
		40	0.0865	0.0093	0.0072	0.0025	0.0014	0.0707
		4	0.123	−0.0231	−0.0128	−0.0175	0.0015	0.0000
		20	0.1131	−0.0126	−0.0069	−0.0115	0.0015	0.0000
3	0.0007	50	0.4977	−0.0048	−0.0034	−0.0027	0.0022	0.2175
		6	0.5447	0.0093	0.0053	0.006	0.0022	0.0064
		5	0.63	0.0039	0.002	0.0027	0.0022	0.2168
		37	0.6767	0.005	0.0029	0.0036	0.0022	0.0973
4	0.72	15	13.11	0.9	0.32	1.03	0.11	0.0000
		2	13.79	1.55	0.55	1.27	0.11	0.0000
		9	14.69	0.62	−0.22	0.5	0.11	0.0000
		7	17.12	1.11	0.25	0.97	0.11	0.0000

for Set 3, the calculated test statistics for $H_0 : D_6 = D_{37}$ is $z_c = 0.7$ and $2P(z \geq 0.7) = 0.479$. Similar tests can be implemented to assess that the analytical methods do not differ by more than a priori fixed acceptance specification.

The estimates of total biases plotted against the averages of two methods are shown in Figs. 1–4, along with their individual 95% confidence intervals. Tolerance limits for 95% confidence and 99% coverage are also calculated. Most of $\widehat{D}_i$ are positive and above the zero line throughout the range in Set 1 and Set 4 indicating that the test method counts WBCs or measures HGB consistently higher than the reference method. Methods agree in 30–70–80% NE range of our sampled subjects because most of $\widehat{D}_i$ are clustered around the zero line (Fig. 3). Most of the $\widehat{D}_i$ are also within the tolerance limits. There is a *CB* in the range of 2% to about 8% MO in Set 2 (Fig. 2). Meanwhile, the obvious downward trend in the range of greater than 10% indicates that the test method counts less MO than the reference. Thus, the two methods do not agree with each other in the upper range.

Sometimes it is valuable to partition the entire range of measurements in specific regions and test differences between biases for each region. This approach may be important when a relatively wide range of measurements is available and *PB* is significant. To illustrate this, we partitioned the entire range of MO in set 2 into two regions, 2%–9% and greater than 9%. Let denote the total biases as D_A and D_B for the first and the second region, respectively. The estimates of total biases for each region are $\widehat{D}_A = -0.0020$ and $\widehat{D}_B = -0.0096$, whereas their standard error are $se(\widehat{D}_A) = 0.0008$ and $se(\widehat{D}_B) = 0.0018$. Both $\widehat{D}_A$ and $\widehat{D}_B$ are statistically

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

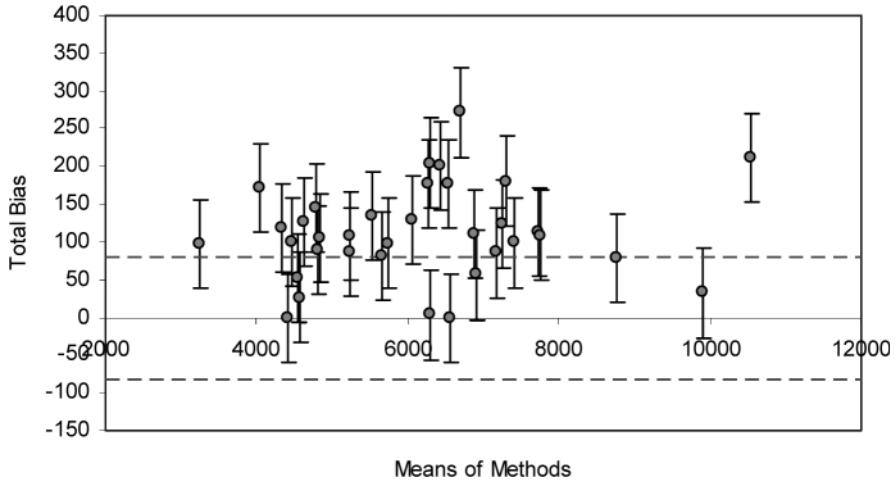

Figure 1. Estimates of total bias between two methods, their confidence intervals, and approximate tolerance limits (···) plotted against the average between two methods of Set 1.

significant (p-values = 0.009 and <0.0001, respectively) as well as statistically different from each other (p-value < 0.0001).

DISCUSSION

A test and a reference method are usually compared on the basis of one of the following hypotheses: the hypothesis of statistical identity within the inherent

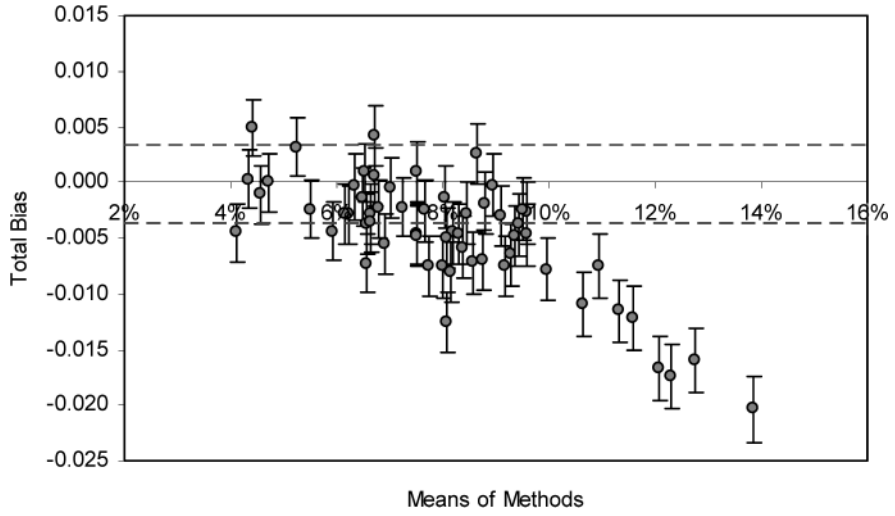

Figure 2. Estimates of total bias between two methods, their 95% confidence intervals, and approximate tolerance limits (···) plotted against the average between two methods of Set 2.

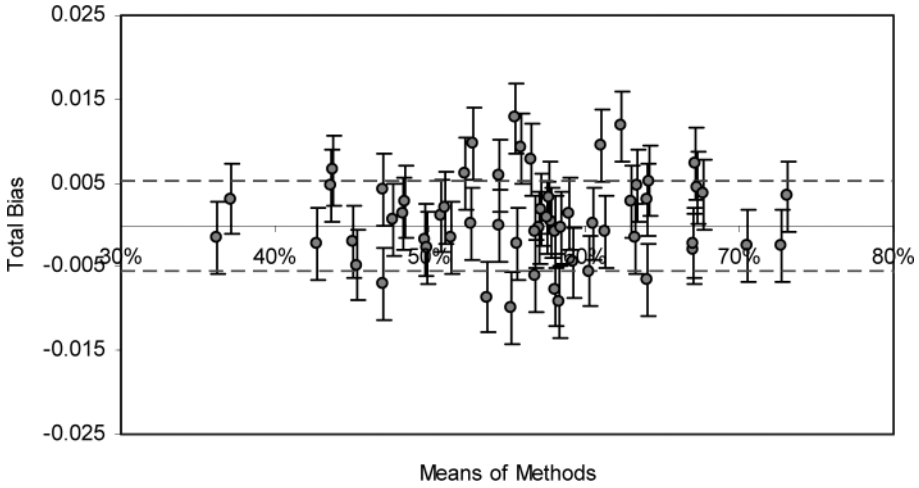

Figure 3. Estimates of total bias between two methods, their 95% confidence intervals, and approximate tolerance limits (···) plotted against the average between two methods of Set 3.

imprecision or the hypothesis of some preset clinical/analytical differences. The first hypothesis depends on the power of the test and sometimes for relatively highly precise measuring devices even small and nonrelevant differences are tested as statistically significant. Although good understanding of what constitutes a relevant difference is needed beforehand to establish sound specification for testing the second hypothesis. Several statistical methods for testing these hypotheses are mentioned in the introduction section. Besides hypothesis testing, graphical interpretation is another popular way for assessing agreement. Altman and Bland (1983) recommended the difference plot, where on the abscissa they used the mean values

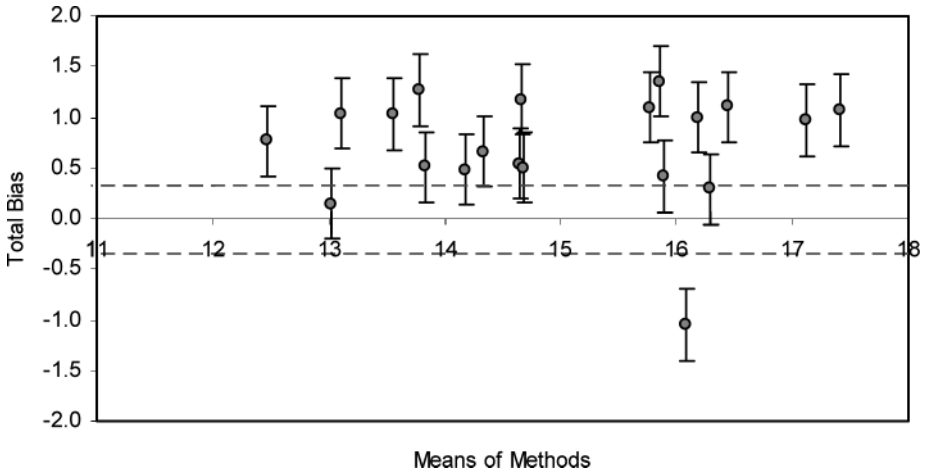

Figure 4. Estimates of total bias between two methods, their 95% confidence intervals, and approximate tolerance limits (···) plotted against the average between two methods of Set 4.

for each subject of the methods to be compared and on the ordinate they plotted the observed differences between methods for the respective subjects. Despite the criticism for lacking an objective criterion for acceptability (Petersen et al., 1997), this approach is found to be very informative and useful in clinical chemistry.

Regardless of the approach used for assessing agreement, the estimation of biases between a test and a reference is of primary interest. These estimates can be either used to test $H_0 : D_i = 0$ or H_0: $D_i \leq sp$ (where sp is a preset specification), and/or in the difference plot. The difference plots presented in Figs. 1–4 are similar to the Altman and Bland (1983) recommendation, but the estimates of total bias instead of the observed differences are plotted on the ordinates. Furthermore, the partitioning of total biases in terms of CB and PB_i gives an insight on the sources of disagreement between two methods and helps designers and manufacturers define a remedial strategy (Magari, 2002). For instance, they may relate the estimates of biases to a meaningful significant difference or to differences on specific ranges if tiered specification are used. Designers may also interpret the $\widehat{D}_i$ vs. the averages scatter according to characteristics such as position, magnitude of the spread, and trend. The position of the scatter is mainly affected by the presence and the value of CB. The $\widehat{D}_i$ are scattered around the zero line when CB is nonsignificant but may be shifted downward or upward when CB is significant. In cases when PB_i are nonsignificant, D_i is a sole function of CB, and subject-to-subject D_i values are not significantly different. Practically, a systematic adjustment (e.g., correction factor) may correct for this bias.

Because CB affects position only, the presence of PB_i in some or in all subjects will affect the magnitude of the spread and/or the trend of the $\widehat{D}_i$ scatter. Significant but non-trending PB_i effects may influence the values of D_i, which will scatter widely in a random fashion throughout the range of measurements. This type of bias may be important for practical purposes if underlying cause factors are identified. The biological characteristics of the subject and random noise factors other then imprecision may affect the measured response by a certain device (Magari, 2002, Houwen, 1990).

The scatter of $\widehat{D}_i$ values is trending when the relative effects of PB_i are related to the range of measurements. This scenario is mostly associated with the lack of linearity of the test method, which is the ability to provide results that are directly proportional to the concentration of the analyte in the sample. However, test and reference may agree in the lower range and show proportional bias in the upper range or vice versa. Or both methods may agree in the center and be biased in the extremes. In conclusions, the approach for biases estimation discussed here is applicable to a wide range of measuring devices used for clinical diagnostic. Ideally, the estimates of biases would assist in identifying potential problems and design corrective actions.

REFERENCES

Altman, D. G., Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician* 32:307–317.

Bartko, J. J. (1994). General methodology II. Measures of agreement: A single procedure. *Stat. Med.* 13:737–745.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Blackwood, L. G., Bradley, E. L. (1991). An omnibus test for comparing two measuring devices. *J. Qual. Tech.* 23:12–16.

Cornbleet, P. J., Gochman, N. (1979). Incorrect least-squares regression coefficient in method-comparison analysis. *Clin. Chem.* 25:432–438.

Golub, G. H. (1973). Some modified matrix eigenvalue problems. *SIAM Rev.* 15:318–334.

Houwen, B. (1990). Random errors in haematology tests: A process control approach. *Clin. Lab. Haematol.* 12:157–168.

Hutson, A. D., Wilson, D. C., Geiser, E. A. (1998). Measuring relative agreement: Echocardiographer versus computer. *J. Agr. Biol. Env. Stat.* 3:163–174.

Jones, A. R., Twedt, D., Swaim, W., Gottfried, E. (1996). Diurnal change of blood count analytes in normal subjects. *Am. J. Clin. Pathol.* 106:723–727.

Krouwer, J. S. (1972). Setting performance goals and evaluating total analytical error for diagnostic assays. *Clin. Chem.* 48:919–927.

Lin, L. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.

Lindstrom, M. J., Bates, D. M. (1998). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *J. Am. Stat. Assoc.* 83:1014–1022.

Linnet, K. (1993). Evaluation of regression procedures for method comparison studies. *Clin. Chem.* 39:424–432.

Magari, R. T. (2002). Evaluating agreement between two analytical methods in clinical chemistry. *Clin. Chem. Lab. Med.* 38:1021–1025.

Passing, H., Bablok, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J. Clin. Chem. Clin. Biochem.* 21:709–720.

Petersen, P. H., Stöckl, D., Blaabjerg, O., Pedersen, B., Birkemose, E., Thienpont, L., Lassen, J. F., Kjeldsen, J. (1997). Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots. *Clin. Chem.* 43:2039–2046.

Pinheiro, J. C., Bates, D. M. (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model. *Comp. Graph. Stat.* 4:12–35.

SAS/Stat User's Guide, Version 8. 2000. SAS Institute, Cary, NC.

St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics* 54:537–545.

Westgard, J. O., Hunt, M. R. (1972). Use and interpretation of common statistical tests in method-comparison studies. *Clin. Chem.* 19:49–57.

Received August 2003
Accepted November 2003

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016