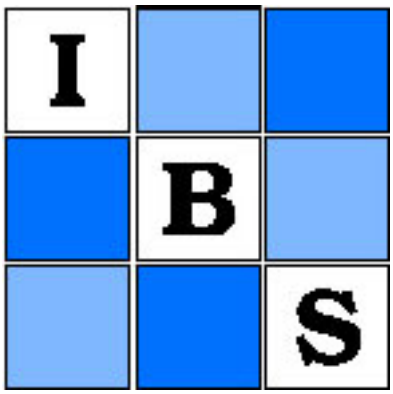

BIOMETRICS 54, 537–545
June 1998

Evaluating Agreement with a Gold Standard in Method Comparison Studies

Roy T. St. Laurent

Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, Arizona 86011-5717, U.S.A.

SUMMARY

We develop a statistical model for method comparison studies where a gold standard is present and propose a measure of agreement. This measure can be interpreted as a population correlation coefficient in a constrained bivariate model. An estimator of this coefficient is proposed and its statistical properties explored. Applications of the new methodology to data from the medical literature are presented.

1. Introduction

Comparison studies are used widely in experimental sciences to assess the degree of agreement between two or more methods of measurement of some quantity. Lewis et al. (1991) distinguish between types of method comparison studies based on whether or not calibration is required and whether or not a known accurate and precise standard of measurement, a gold standard, is available. They define three types of such studies: (1) calibration problems, in which an approximate method of measurement is to be calibrated to a gold standard; (2) conversion problems, in which two approximate methods not measured on the same scale need to be calibrated one to the other; and (3) [approximation] comparison problems, in which two approximate methods measured on the same scale are to be compared in order to assess the extent to which they agree.

Missing from this typography is a fourth type of method comparison problem studied here: (4) gold-standard comparison problems, in which an approximate method of measurement is compared to a gold standard in order to assess the degree to which the approximate measure agrees with the gold standard. This problem differs from calibration and conversion problems in that it is assumed that the methods being compared yield measurements on the same scale so that no calibration is desired, and it differs from conversion and approximation comparison problems in that a gold standard is involved.

Comparison problems where a gold standard is present are common in the literature, as the following examples illustrate.

Example I (*Prigent et al.*, 1991): The seriousness of myocardial infarction may be related to the percentage of heart muscle mass affected. This percentage can be accurately determined by pathologic examination. To assess and compare the ability of single photon emission computed tomography (SPECT) and planar myocardial perfusion imaging methods to noninvasively determine the percentage of heart muscle mass affected, investigators induced myocardial infarcts in 12 dogs and, in each, measured infarct size by 3 methods: SPECT, planar imaging, and pathologic examination. Investigators wish to determine the extent to which SPECT and planar imaging measurements agree with the pathologic measurements and whether or not one method (SPECT or planar imaging) matches pathology better than the other.

Example II (*Raz, Chenevert, and Fernandez*, 1994): Investigators compare Lorentz and spline models of the effect of inhomogeneity in a magnetic field applied to estimation of spin-spin relaxation

Email address: Roy.St.Laurent@nau.edu

Key words: Agreement; Correlation; Gold standard; Intraclass correlation coefficient; Method comparison study; Random effects model.

time using a model of spin echoes acquired from six chemical phantoms. Under both the Lorentz and spline models, single spin echo and multiecho estimates of the spin-spin relaxation time are available. Multiecho estimates from the spline model are treated as a gold standard, and the extent to which single-echo estimates calculated from the Lorentz or spline models agree with the multiecho estimates is evaluated. This process is repeated with multiecho estimates from the Lorentz model treated as a gold standard.

Example III: Wax, Hoffman, and Goldfrank (1992) assess agreement of blood alcohol concentration (BAC) as measured by a rapid response electrochemical meter, with gold standard BAC determined by blood immunoassay.

Examples in statistics and entomology of comparison studies where a gold standard is present may be found in St. Laurent and Gebremariam (1993) and Dosdall, Herbut, and Cowle (1994), respectively. Additional medical examples include Christoffersson et al. (1987), de Yang et al. (1991), Finkelstein et al. (1987), Lin (1989), and Yamagishi et al. (1991).

Assessment of agreement between two or more approximate methods has been studied at length. A standard model (Fleiss, 1986, Chapter 1; Lord and Novick, 1968, Chapter 3; Donner, 1986) used widely in psychology and clinical medicine for measuring reproducibility of measurements (agreement) is the one-way random-effects model

$$X_{ij} = T_i + \varepsilon_{ij}, \tag{1}$$

where for the ith unit (or subject), $i = 1, \ldots, n$, X_{ij} is the measurement made by the jth method, $j = 1, \ldots, k$, T_i is a random variable with mean μ and variance σ_T^2, and ε_{ij} is an unobservable measurement error independent of T_i, with mean 0 and variance σ^2. For $k = 2$, values for (X_{i1}, X_{i2}) are observed, and T_i is the unobserved true value associated with the ith unit.

Typically, agreement between X_{i1} and X_{i2} is measured by the intraclass correlation coefficient $\rho_I = \sigma_T^2/(\sigma_T^2 + \sigma^2)$, where ρ_I is the correlation between X_{i1} and X_{i2}.

Under the assumption that the random variables T_i and ε_{ij} are normally distributed, the maximum likelihood estimator of ρ_I is $r_m = 1/(1 + S_d^2/C_m)$, where $S_d^2 = \Sigma(X_{i1} - X_{i2})^2$, $C_m = 2S_{12} - n(\bar{X}_1 - \bar{X}_2)^2/2$, $\bar{X}_j$ is the mean of measurements on the jth method, and $S_{12} = \Sigma(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$. The analysis-of-variance-based estimator of ρ_I may be written similarly as $r_a = 1/(1 + S_d^2/C_a)$ for appropriate choice of scale factor C_a. Donner (1986) provides an overview of estimators and inference procedures for the intraclass correlation coefficient for the generalization of model (1) allowing n_j observations per measurement method.

Starting from a bivariate normal model, Lin (1989) develops a concordance correlation coefficient similar to r_m, $r_c = 1/(1 + S_d^2/C_c)$, where $C_c = 2S_{12}$. He proposes that inference for ρ_I be based on the asymptotic distribution of a transform of r_c. Chinchilli et al. (1996) extend Lin's results to repeated measurements designs and use a nonparametric bootstrap to construct confidence intervals.

In what follows, model (1) is modified for use in a gold-standard comparison problem and a gold-standard correlation ρ is developed (Section 2). In Section 3, an estimator of ρ is proposed and several approaches to inference are briefly considered. In Section 4, we generalize the results to consider the comparison of $J \geq 2$ approximate methods of measurement to a gold-standard method, and finally, in Section 5, we apply the methodology to several examples. Proofs of key technical results are given in the Appendix.

2. Modeling Agreement Between an Approximate Method and a Gold Standard

To assess agreement of an approximate method of measurement with a gold standard, consider the model

$$X_i = G_i + \varepsilon_i, \tag{2}$$

where X_i is the approximate measurement on the ith unit, $i = 1, \ldots, n$; G_i is the corresponding gold-standard measurement, a random variable with mean μ and variance σ_G^2; and ε_i is a measurement error associated with the approximate measurement, independent of G_i, with mean 0 and variance σ^2. This is essentially the random-effects model (1) for $j = 1$, except that the random effect G_i is observed. The random effect G_i accounts for the unit-to-unit variability in the population of gold-standard measurements. Model (2) says that for each unit the approximate measurement would be in perfect agreement with the corresponding gold standard were it not for an additive measurement error.

From (2), $\text{cov}(X_i, G_i) = \text{var}(G_i) = \sigma_G^2$ and $\text{var}(X_i) = \sigma_G^2 + \sigma^2$, and it follows that $\rho = \sigma_G^2/(\sigma_G^2 + \sigma^2)$ is the square of the correlation between X_i and G_i, identical in form to the intraclass correlation ρ_I. Thus, ρ itself, or $\rho^{1/2}$, the correlation between X_i and G_i, can be used to measure agreement between approximate and gold standard. Using ρ (or $\rho^{1/2}$) to measure agreement means that agreement is evaluated relative to the variability in the population of gold-standard measurements (σ_G^2). For a fixed measurement error variability σ^2, ρ is an increasing function of σ_G^2. Thus, ρ is a relative measure of agreement dependent on σ_G^2. Any notion of what constitutes agreement in an absolute sense requires subject-matter specification as to what is an acceptable upper bound for σ^2 beyond which—regardless of the value of σ_G^2 or ρ—one would conclude that the approximate and gold standards were not in sufficient agreement. Altman and Bland (1983) address this issue in the context of assessment of agreement between two approximate methods of measurement.

Model (2) is equivalent to a bivariate model for independent identically distributed random vectors $\mathbf{U}_i = (X_i, G_i)^{\mathrm{T}}, i = 1, \ldots, n$, where $\text{E}(\mathbf{U}_i) = (\mu, \mu)^{\mathrm{T}}$ and

$$\text{cov}(\mathbf{U}_i) = \begin{pmatrix} \tau^2 & \delta^2 \\ \delta^2 & \delta^2 \end{pmatrix} \tag{3}$$

and where $\delta^2 \leq \tau^2$ in order for (3) to be nonnegative definite. The correspondence with model (2) is $\tau^2 = \sigma_G^2 + \sigma^2$ and $\delta^2 = \sigma_G^2$.

For a general bivariate model $\mathbf{U} = (X, G)^{\mathrm{T}}$, with arbitrary mean vector and nonsingular covariance matrix, model (2) may be characterized as the bivariate model for which $\min_{a,b \in \mathbb{R}} \text{E}[X - (a + bG)]^2$ is achieved at $a = 0$ and $b = 1$, i.e., the model for which the linear regression of X on G is the line of agreement $X = G$. Conversely, if for a bivariate model the linear regression is the line of agreement, then the bivariate model is equivalent to model (2).

This relationship between model (2) and the line of agreement is consistent with the intuitive notion that, if an approximate measure is in perfect agreement with a gold standard, then the points in a plot of the approximate measure vs. the gold standard will fall on the line of agreement $X = G$.

3. Estimating Agreement When a Gold Standard Is Present

3.1 *Point Estimation*

We consider four *ad hoc* estimators of $\rho = \sigma_G^2/(\sigma_G^2 + \sigma^2)$ based on $D_i = X_i - G_i$, the observed difference between the measurements on the approximate and gold-standard methods, and the sums of squares $S_{XX} = \Sigma\,(X_i - \bar{X})^2$, $S_{GG} = \Sigma\,(G_i - \bar{G})^2$, and $S_{DD} = \Sigma\,D_i^2$, where $\bar{X} = \Sigma\,X_i/n$ and $\bar{G} = \Sigma\,G_i/n$.

(i) Since $\rho^{1/2}$ is the correlation between X and G, estimate ρ by r_p^2, the square of the sample Pearson correlation coefficient between X_i and G_i. However, r_p^2 does not distinguish situations where X and G strongly agree (high correlation about the line of agreement) from those where a strong linear relationship is apparent but agreement is not, and therefore r_p^2 would be a poor estimator of ρ.

(ii) Model (2) is equivalent to a bivariate model in which the regression $\text{E}(X \mid G = g)$ is constrained. Since ρ may be thought of as the proportion of variation in X explained by the (constrained) linear relationship between X and G, estimate ρ by $1 - S_{DD}/S_{XX}$, the proportion of variation explained by the agreement between X_i and $G_i, i = 1, \ldots, n$, relative to the variability in the n sampled values of X about their mean, a statistic similar to the coefficient of determination in linear regression.

This approach treats the gold-standard measurements as fixed rather than as a random sample from a population. As the models being compared, $\text{E}(X_i \mid G_i = g_i) = \mu$ and $\text{E}(X_i \mid G_i = g_i) = g_i$, are not nested, there is the added difficulty of interpreting $1 - S_{DD}/S_{XX}$ as the proportion of variation explained by agreement, and there is no guarantee that $1 - S_{DD}/S_{XX}$ will be nonnegative.

(iii) Estimate the numerator and denominator of ρ separately. $S_{GG}/(n-1)$ is an unbiased estimator of σ_G^2 and $S_{XX}/(n-1)$ is an unbiased estimator of $\sigma_G^2 + \sigma^2$, suggesting S_{GG}/S_{XX} as an estimator of ρ. However, while $0 \leq \rho \leq 1$, the ratio S_{GG}/S_{XX} may exceed 1.

(iv) Estimate the variance components separately. As noted in (iii), $S_{GG}/(n-1)$ is an unbiased estimator of σ_G^2. Further, S_{DD}/n is an unbiased estimator of σ^2 since, under the assumptions of model (2), $\text{E}(D_i) = 0$ and $\text{var}(D_i) = \text{var}(\varepsilon_i) = \sigma^2$. This suggests

$$1/[1 + (n-1)S_{DD}/(nS_{GG})]$$

as a possible estimator of $\rho = 1/[1 + \sigma^2/\sigma_G^2]$ or the asymptotically equivalent estimator

$$r_g^2 = 1/[1 + S_{DD}/S_{GG}].$$

Both these estimators have the same range as ρ and are similar in form to the estimators of ρ_I discussed in Section 1.

Further support for the choice of r_g^2 as the preferred estimator of ρ is provided by the result that, if ε_i and G_i are normally distributed, then the maximum likelihood estimator of ρ is r_g^2. This is a special case of a more general result presented as Proposition 3 in Section 4.

3.2 *Properties of* r_g^2

As mentioned in (i) above, the sample correlation coefficient r_p measures strength of linear relationship, not agreement. However, r_p^2 and the estimated gold-standard correlation coefficient r_g^2 are closely related. In the Appendix, we show that

$$1/r_p^2 - 1 = RSS/(\hat{\beta}^2 S_{GG}), \tag{4}$$

where $\hat{\beta}$ is the least squares estimator of the slope in the unconstrained linear regression of X on G and RSS is the corresponding residual sum of squares. For the gold-standard correlation, we have

$$1/r_g^2 - 1 = \frac{RSS + \sum(\hat{X}_i - G_i)^2}{S_{GG}}, \tag{5}$$

where $\hat{X}_i$ is the ith fitted value from the unconstrained fit, so that $\Sigma\,(\hat{X}_i - G_i)^2$ is a measure of the discrepancy between the values predicted for X by the unconstrained least squares fit and the values predicted by the line of agreement.

If the unconstrained fitted regression line coincides with the line of agreement, then $\hat{\beta} = 1$ and $\Sigma\,(\hat{X}_i - G_i)^2 = 0$, in which case $r_p^2 = r_g^2$. An advantage r_g^2 has over r_p^2 is that, when model (2) does not hold exactly, an extra penalty may be incurred: r_g^2 measures not only the strength of agreement but also the extent to which the model of agreement fits the data. Since even when $RSS = 0$ and hence the unconstrained line provides a perfect fit to the data so that $r_p^2 = 1$, the dependency of r_g^2 on $\Sigma\,(\hat{X}_i - G_i)^2$ implies that r_g^2 will be less than 1.

Examination of the form of the usual F statistic used in regression to compare the unconstrained fit to the fit of the model of agreement clarifies the importance of RSS and $\Sigma\,(\hat{X}_i - G_i)^2$ in this problem: $F = [(n-2)/2]\,\Sigma\,(\hat{X}_i - G_i)^2/RSS$. In the regression setting, the model of agreement is rejected when $\Sigma\,(\hat{X}_i - G_i)^2$ is large relative to RSS.

It is apparent from (4) and (5) that, if the fitted regression line of X on G has slope $\hat{\beta} = 1$, then $r_g^2 \leq r_p^2$ with equality if and only if the estimated intercept $\hat{\alpha} = 0$. However, this inequality does not hold uniformly for other values of $(\hat{\alpha}, \hat{\beta})$. In the Appendix, we show that r_g^2 and r_p^2 are related by

$$1/r_g^2 = \hat{\beta}^2/r_p^2 - 2(\hat{\beta} - 1) + (n/S_{GG})[\hat{\alpha} + (\hat{\beta} - 1)\bar{G}]^2. \tag{6}$$

Noting that the last term on the right-hand side of (6) is nonnegative and that the remaining terms there collectively are a quadratic, convex function of $\hat{\beta}$ for each value of r_p^2, it may be shown that

[1] if $\hat{\beta} \geq 1$, then $r_g^2 \leq r_p^2$, with equality if and only if $\hat{\beta} = 1$ and $\bar{X} = \bar{G}$;
[2] if $2r_p^2 - 1 < \hat{\beta} < 1$, then $r_g^2 \leq 1/(2 - r_p^2)$, with equality if and only if $\hat{\beta} = r_p^2$ and $\bar{X} = \bar{G}$; and
[3] if $\hat{\beta} \leq 2r_p^2 - 1$, then $r_g^2 \leq r_p^2$ with equality if and only if $\hat{\beta} = 2r_p^2 - 1$ and $\bar{X} = \bar{G}$.

While the inequality $r_g^2 \leq r_p^2$ does not hold when $2r_p^2 - 1 < \hat{\beta} < 1$, the upper bound $1/(2 - r_p^2)$ in [2] is not that much greater than r_p^2 in those situations where one might reasonably wish to assess agreement. For example, for any r_p^2 greater than 0.75, $r_g^2 \leq r_p^2 + 0.05$.

What are the consequences of applying Lin's concordance correlation coefficient r_c when a gold standard is present? It can easily be shown that, for $\hat{\beta} \neq 0$,

$$1/r_g^2 - 1 = 2\hat{\beta}(1/r_c - 1).$$

It follows that r_c underestimates agreement when $r_g \leq 1/(2\hat{\beta} - 1)$, i.e., $r_c < r_g$, and overestimates agreement otherwise. In particular, for $\hat{\beta} \leq 1$, $r_c < r_g$. For $\hat{\beta}$ close to 1, the maximum discrepancy is largest when $\hat{\beta} < 1$, e.g., when $9/10 \leq \hat{\beta} \leq 10/9$, $\max|r_c - r_g|$ occurs when $\hat{\beta} = 9/10$, in which case, to two decimal places, $\max|r_c - r_g| = r_g - r_c = 0.33 - 0.18 = 0.15$.

3.3 *Inference Concerning* ρ

PROPOSITION 1: *If in addition to the assumptions of* (2), *the random variables* ε_i *and* G_i *are normally distributed, then the statistic* $(1-1/n)(1/r_g^2-1)$ *is distributed as* $[1/\rho-1]F_{n,n-1}$, *where* $F_{n,n-1}$ *is an* F *random variable with* n *and* $n-1$ *degrees of freedom.*

For problems in which the normal distribution assumptions are reasonable, this result can serve as the basis for testing hypotheses and constructing confidence intervals for ρ. For example, to construct a $100(1-\xi)$ percent lower confidence limit for ρ for a given ξ, find F_L the critical value from the F distribution with n and $n-1$ degrees of freedom such that $P\{F_{n,n-1} < F_L\} = \xi$, then the set of all ρ that satisfy

$$\frac{F_L}{F_L + [1/r_g^2 - 1](n-1)/n} \leq \rho \leq 1$$

is a one-sided confidence interval for ρ.

PROPOSITION 2: *If the fourth central moments of the distributions of* ε_i *and* G_i *are both finite* (μ_4, μ_{4G}, *respectively*), *then as* $n \to \infty$, *the distribution of* $\ln(1/r_g^2 - 1)$ *converges to a normal with mean* $\ln(1/\rho - 1)$ *and asymptotic variance* $\gamma_{AV}^2 = (\mu_{4G}/\sigma_G^4 + \mu_4/\sigma^4 - 2)/n$.

The nonparametric bootstrap provides an alternative approach to statistical inference in those circumstances where a full specification of the distribution of ε_i and G_i is not available. In particular, the bias-corrected accelerated percentile (BC_a) confidence interval method is competitive and, in many cases, superior to standard asymptotic normal-based intervals (Efron and Tibshirani, 1993, Chapter 22). The method also has the desirable property that the BC_a interval for a monotonic transformation of the parameter θ, say $\phi = f(\theta)$ based on $\hat{\phi} = f(\hat{\theta})$, is identical to the interval formed by transforming the endpoints of the BC_a interval for θ.

4. Several Approximate Measures and a Gold Standard

In many applications (including Examples I and II of Section 1), there is more than one approximate method of measurement to compare to the gold standard. Suppose that $J \geq 2$ approximate methods of measurement, indexed by $j = 1, \ldots, J$, are to be compared to a gold standard. For simplicity, assume that each unit $i = 1, \ldots, n$ is measured by all J approximate methods and the gold standard, resulting in n $(J+1)$-tuples $(X_{i1}, \ldots, X_{iJ}, G_i)$. Model the relationship between X_{ij} and G_i as

$$X_{ij} = G_i + \varepsilon_{ij}, \tag{7}$$

where ε_{ij} is the measurement error on the ith unit by the jth approximate method and we take $\mathrm{E}(G_i) = \mu$, $\mathrm{E}(\varepsilon_{ij}) = 0, \mathrm{var}(G_i) = \sigma_G^2$. In addition, assume that G_i, the gold-standard measurement on the ith unit, is independent of $\varepsilon = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})^{\mathrm{T}}$, the vector of measurement errors. To allow for the possibility that measurement errors on the ith unit are correlated across approximate methods, assume a general covariance structure on ε, $\mathrm{cov}(\varepsilon) = \boldsymbol{\Sigma}$, for some positive definite $J \times J$ matrix $\boldsymbol{\Sigma}$ with (j,k)th element σ_{jk}.

A consequence of these assumptions is that $\mathrm{cov}(X_{ij}, G_i) = \sigma_G^2$ and $\mathrm{cov}(X_{ij}, X_{ik}) = \sigma_G^2 + \sigma_{jk}$. Thus, $\rho_j = \sigma_G^2/(\sigma_G^2 + \sigma_{jj})$ is the square of the correlation between X_{ij} and G_i. Note that the ρ_j, $j = 1, \ldots, J$, do not depend on the off-diagonal elements of $\boldsymbol{\Sigma}$. Since there is no requirement that $\sigma_{jj} = \sigma_{kk}$, this model allows for different levels of agreement ρ_j between each approximate method and the gold standard.

PROPOSITION 3: *If, in addition to the assumptions of* (7), *the random variables* G_i *and* ε_{ij} *are normally distributed, then the maximum likelihood estimator of* ρ_j *is* $r_{g(j)}^2 = 1/[1 + S_{DD(j)}/S_{GG}]$, *where* $S_{DD(j)} = \Sigma\,(X_{ij} - G_i)^2$ *for* $j = 1, \ldots, J$.

Since model (7) reduces to model (2) when $J = 1$, marginal inference about the agreement ρ_j between a single approximate method and the gold standard can proceed as described in Section 3.3.

When two or more approximate methods are available, a natural question to ask is whether one is in closer agreement with the gold standard than the others. Here we restrict our attention to the case $J = 2$. A test of the null hypothesis that the level of agreement of the approximate methods with the gold standard is the same, vs. the alternative hypothesis that the level of agreement differs across the approximate methods, corresponds to testing H_0: $\rho_1 = \rho_2$ vs. H_a: $\rho_1 \neq \rho_2$. Since $\rho_1 = \rho_2$ if and only if $\sigma_{11} = \sigma_{22}$, this is equivalent to testing H_0: $\sigma_{11} = \sigma_{22}$ vs. H_a: $\sigma_{11} \neq \sigma_{22}$. Under the assumptions of model (7), for fixed j, the D_{ij} are independent with $\mathrm{E}(D_{ij}) = 0$ and $\mathrm{var}(D_{ij}) = \sigma_{jj}$,

$i = 1, \ldots, n$. However, $\text{cov}(D_{i1}, D_{i2}) = \sigma_{12}$. Thus, the problem of testing for equal agreement between the two approximate methods is equivalent to testing for equal variances among correlated observations, each with mean 0, based on samples of size n from each population.

For the slightly more general problem where $\text{E}(D_{ij}) = \mu_j$, the hypothesis test H_0: $\sigma_{11} = \sigma_{22}$ vs. H_a: $\sigma_{11} \neq \sigma_{22}$ has been considered by a number of authors. For the normal case, Pitman (1939) and Morgan (1939) derive a test based on the correlation between $D_{i1} + D_{i2}$ and $D_{i1} - D_{i2}$. Grambsch (1994) reviews a number of proposed test statistics for the nonnormal case and investigates the robustness of their asymptotic normality under various distributional assumptions.

The approach used here is to construct intervals for the difference $\rho_1^{1/2} - \rho_2^{1/2}$ using the nonparametric bootstrap BC_a interval discussed at the end of Section 3.3.

5. Examples

Example I (*myocardial infarct size*): Prigent et al. (1991) plot the percentage of heart muscle mass affected by an induced myocardial infarct as measured by (corrected) SPECT (vertical axis) vs. that measured by pathology (gold standard, horizontal axis) for $n = 12$ dogs. The plot shows that the line of agreement provides a good approximation to the least squares fitted line ($\hat{\alpha} = 0.6$, $\hat{\beta} = 0.90$, and the p-value for testing $(\alpha, \beta) = (0, 1)$ is reported as 0.48). The investigators summarize the extent of agreement between SPECT and pathology by calculating the Pearson correlation coefficient $r_p = 0.85$ and Lin's concordance correlation coefficient $r_c = 0.84$. Though the authors do not report the raw data, based on the relationship $1/r_g^2 - 1 = 2\hat{\beta}(1/r_c - 1)$, we can calculate the gold-standard correlation coefficient to be $r_g = 0.86$. This value is quite similar to both r_p and r_c, which is to be expected since the regression line is essentially the line of agreement.

Agreement between (corrected) planar imaging and pathology is not as good. The scatterplot of paired values shows clear departure from the line of agreement ($\hat{\alpha} = 7$, $\hat{\beta} = 1.13$, $F = 6.84$, p-value < 0.05). Here $r_p = 0.75$, $r_c = 0.49$, and we obtain $r_g = 0.55$. The discrepancy between r_g and r_p may be attributed to the fact that the linear regression is substantially different from the line of agreement.

The SPECT measurements are much closer in agreement ($r_g = 0.86$) with pathology than measurements obtained by planar imaging ($r_g = 0.55$). Assuming normality and applying a Pitman-like test for equal variances of D_{i1} and D_{i2} (equivalent to testing $\rho_1 = \rho_2$), we obtain a statistic that, under the null hypothesis, is distributed as a Student's t on $n - 1$ degrees of freedom. For this data, $n - 1 = 11$, and it can be shown that $t \geq 3.63$. Therefore, the p-value for a two-sided test is less than 0.004, confirming the suspicion that SPECT is superior to planar imaging in its agreement with the pathologic measurement.

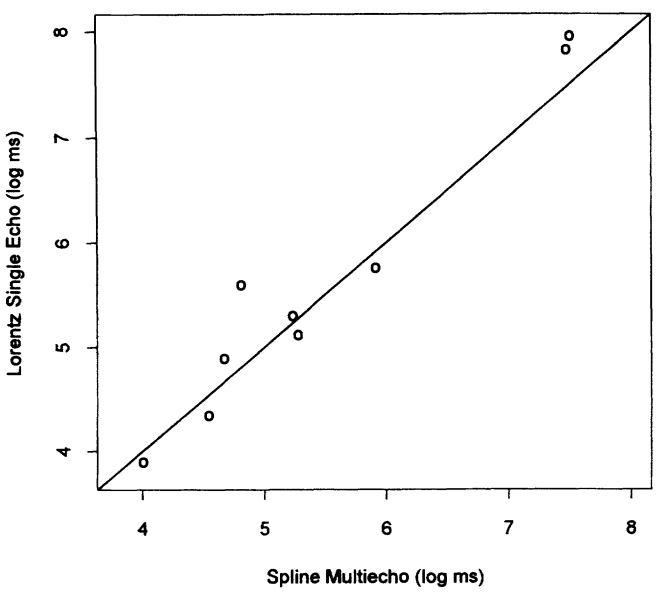

Figure 1. Plot of Lorentz model single-echo estimates of spin-spin relaxation time vs. spline model multiecho estimates and the line of agreement. Scale on both axes is natural log of milliseconds.

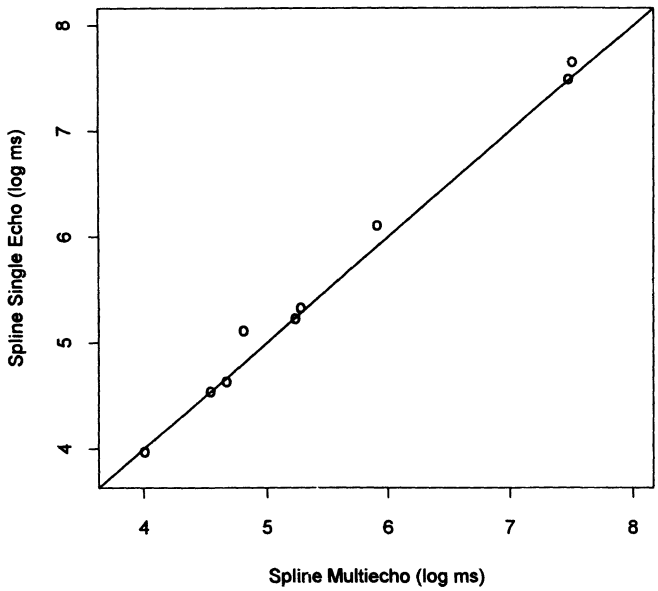

Figure 2. Plot of spline model single-echo estimates of spin-spin relaxation time vs. spline model multiecho estimates and the line of agreement. Scale on both axes is natural log milliseconds.

Example II (*spin-spin relaxation times*): We concentrate here only on the portion of the analysis that treats the spline multiecho estimates of spin-spin relaxation time as a gold standard. A plot of the $n = 9$ Lorentz single-echo estimates vs. the spline multiecho estimates is given in Figure 1. The scale on both axes is natural log milliseconds. Six chemical phantoms were used, three of which had two components each, for a total of nine points. In the original data set of Raz et al. (1994), multiple values were reported for each of the nine phantom/component combinations. To simplify the analysis here, these multiple values have been averaged.

Nonparametric bootstrap BC_a intervals were constructed for each gold-standard correlation separately and also for the difference in the correlations. We took $B = 2000$ bootstrap iterations, sampling with replacement from the triple (X_{i1}, X_{i2}, G_i), $i = 1, \ldots, 9$. Calculations were performed using the S-PLUS software routines provided in Efron and Tibshirani (1993).

The estimated agreement between the Lorentz single-echo estimates and the spline multiecho estimates is $r_g = 0.956$. The two-sided 95% BC_a confidence interval for $\rho^{1/2}$ is $(0.766, 0.984)$.

The spline single-echo estimates are plotted vs. the gold standard in Figure 2. It is apparent from the plot that the agreement is better here than in Figure 1. We have $r_g = 0.992$ and 95% BC_a interval $(0.947, 0.999)$.

While the 95% intervals for the Lorentz and the spline estimates of agreement overlap, it is clear that the Lorentz and spline single-echo estimates are not independent. A 95% BC_a confidence interval for the difference $\rho^{1/2}_{spline} - \rho^{1/2}_{Lorentz}$ is $(0.011, 0.202)$, indicating that there is evidence that the spline single-echo estimates are in better agreement with the multiecho gold standard than are the Lorentz single-echo estimates. While the lower endpoint of this 95% confidence interval is close to zero, it is worth noting that the averaging used to construct the dataset analyzed here may well have had the effect of dampening the differences between the levels of agreement of the two single-echo estimates with the gold standard.

ACKNOWLEDGEMENTS

I thank Jonathan Raz, Ian Harris, and two anonymous referees for helpful comments on earlier drafts of this paper.

RÉSUMÉ

Nous développons un modèle pour des études de comparaison de méthodes en présence d'un standard de référence et proposons une mesure d'agrément. Cette mesure peut être interprétée comme un coefficient de corrélation dans un modèle bivarié avec contrainte. Un estimateur de ce coefficient

est proposé et ces propriétés statistiques étudiées. Des applications de cette nouvelle méthodologie à des données médicales sont présentées.

REFERENCES

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

Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement design. *Biometrics* **52,** 341–353.

Christoffersson, J. O., Welinder, H., Spång, G., Mattsson, S., and Skerfving, S. (1987). Cadmium concentration in the kidney cortex of occupationally exposed workers measured *in vivo* using X-ray fluorescence analysis. *Environmental Research* **42,** 489–499.

de Yang, L., Bairey, C. N., Berman, D. S., Nichols, K. J., Odom-Maryon, T., and Rozanski, A. (1991). Accuracy and reproducibility of left ventricular ejection fraction measurements using an ambulatory radionuclide left ventricular function monitor. *Journal of Nuclear Medicine* **32,** 796–802.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54,** 67–82.

Dosdall, L. M., Herbut, M. J., and Cowle, N. T. (1994). Susceptibilities of species and cultivars of canola and mustard to infestation by root maggots (*Delia* spp.) (Diptera: Anthomyiidae). *The Canadian Entomologist* **126,** 251–260.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Finkelstein, S. M., Kujawa, S. J., Budd, J. R., and Warwick, W. J. (1987). An evaluation of incentive inspirometers for following pulmonary function by self-measurement in the home. *IEEE Transactions on Biomedical Engineering* **BME-34,** 212–216.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments.* New York: Wiley.

Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika* **81,** 359–372.

Lewis, P. A., Jones, P. W., Pollak, J. W., and Tillotson, H. T. (1991). The problem of conversion in method comparison studies. *Applied Statistics* **40,** 105–112.

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

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores,* with contributions by A. Birnbaum. Reading, Massachusetts: Addison-Wesley.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika* **31,** 13–19.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika* **31,** 9–12.

Prigent, F. M., Maddahi, J., Van Train, K. F., and Berman, D. S. (1991). Comparison of thallium-201 SPECT and planar imaging methods for quantification of experimental myocardial infarct size. *American Heart Journal* **122,** 972–979.

Raz, J., Chenevert, T., and Fernandez, E. J. (1994). A flexible spline model of the spin echo with applications to estimation of the spin-spin relaxation time. *Journal of Magnetic Resonance A* **111,** 137–149.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

St. Laurent, R. T. and Gebremariam, A. (1993). Approximate measures of influence in nonlinear regression: A case study. *American Statistical Association 1993 Proceedings of the Statistical Computing Section*, 159–164.

Wax, P. M., Hoffman, R. S., and Goldfrank, L. R. (1992). Rapid quantitative determination of blood alcohol concentration in the emergency department using an electrochemical method. *Annals of Emergency Medicine* **21,** 254–259.

Weisberg, S. (1985). *Applied Linear Regression,* 2nd edition. New York: Wiley.

Yamagishi, M., Hotta, D., Tamai, J., Nakatani, S., and Miyatake, K. (1992). Validity of catheter-tip Doppler technique in assessment of coronary flow velocity and application of spectrum analysis method. *The American Journal of Cardiology* **67,** 758–762.

Received October 1995; *revised June* 1997; *accepted July* 1997.

APPENDIX

Derivation of equation (4). In the unconstrained linear regression of X on G, r_p^2 is identical to the coefficient of determination, $r_p^2 = R^2 = 1 - RSS/S_{XX} = (S_{XX} - RSS)/S_{XX}$. From Weisberg (1985, p. 12), we know that $RSS = S_{XX} - \hat{\beta}^2 S_{GG}$. Solving for S_{XX} gives $S_{XX} = RSS + \hat{\beta}^2 S_{GG}$. Substituting in the expression for r_p^2 above gives $r_p^2 = \hat{\beta}^2 S_{GG}/(RSS + \hat{\beta}^2 S_{GG})$, so that $1/r_p^2 - 1 = RSS/(\hat{\beta}^2 S_{GG})$, as was to be shown.

Derivation of equation (5). From the definition of r_g^2 in Section 3.1, it follows that $1/r_g^2 - 1 = S_{DD}/S_{GG}$. Substituting $(X_i - \hat{X}_i) + (\hat{X}_i - G_i)$ for $X_i - G_i$ in $S_{DD} = \Sigma (X_i - G_i)^2$ and expanding the result gives

$$S_{DD} = \sum \hat{e}_i^2 + \sum (\hat{X}_i - G_i)^2 + 2 \sum \hat{e}_i (\hat{X}_i - G_i),$$

where $\hat{e}_i = (X_i - \hat{X}_i)$ is the ith residual. The first term on the right-hand side of the displayed equation is RSS. The third term is zero since the vector of residuals is orthogonal to the vector of predicted values, i.e., $\Sigma \hat{e}_i \hat{X}_i = 0$; and $\Sigma \hat{e}_i G_i$ is the right-hand side of a normal equation (evaluated at the least squares estimates) for the unconstrained model, hence $\Sigma \hat{e}_i G_i = 0$. Thus, $S_{DD} = RSS + \Sigma (\hat{X}_i - G_i)$, from which it follows that $1/r_g^2 - 1 = [RSS + \Sigma (\hat{X}_i - G_i)^2]/S_{GG}$.

Derivation of equation (6). From (5), $1/r_g^2 = 1 + [RSS + \Sigma (\hat{X}_i - G_i)^2]/S_{GG}$. Solving (4) for RSS gives $RSS = [1/r_p^2 - 1]\hat{\beta}^2 S_{GG}$. Substituting this into the expression for $1/r_g^2$ above and a little algebra gives

$$1/r_g^2 = \hat{\beta}^2/r_p^2 - \hat{\beta}^2 + 1 + \sum (\hat{X}_i - G_i)^2/S_{GG}.$$

Substituting $\hat{X}_i = \hat{\alpha} + \hat{\beta} G_i$ into $\Sigma (\hat{X}_i - G_i)^2/S_{GG}$ and application of additional algebra yields $\Sigma (\hat{X}_i - G_i)^2/S_{GG} = (n/S_{GG})[\hat{\alpha} + (\hat{\beta} - 1)\bar{G}]^2 + (\hat{\beta} - 1)^2$. Substituting this into the displayed expression for $1/r_g^2$ gives the result stated in (6).

Proof of Proposition 1. From the model (2) assumptions and the additional normality assumption, $D_i = X_i - G_i = \varepsilon_i$ and G_i are uncorrelated normal random variables and hence are independent. Note that $\mathrm{E}(D_i) = 0$, $\mathrm{E}(G_i) = \mu$, $\mathrm{var}(D_i) = \sigma^2$, and $\mathrm{var}(G_i) = \sigma_G^2$. Applying standard results for normal random variables, this implies that $S_{DD} = \Sigma D_i^2$ is distributed as $\sigma^2 \chi^2(n)$, where $\chi^2(n)$ is a chi-square random variable with n degrees of freedom; $S_{GG} = \Sigma (G_i - \bar{G})^2$ is distributed as $(\sigma_G^2)\chi^2(n-1)$; and S_{DD} and S_{GG} are independent. Therefore, $(n-1)S_{DD}/(nS_{GG})$ is distributed as $(\sigma^2/\sigma_G^2)F_{n,n-1}$. Noting that $\sigma^2/\sigma_G^2 = 1/\rho - 1$, the result follows.

Proof of Proposition 2. Note that $\ln(1/r_g^2 - 1) = \ln(S_{DD}/n) - \ln(S_{GG}/n)$. The components of the vector $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2)^{\mathrm{T}} = (S_{DD}/n, S_{GG}/n)^{\mathrm{T}}$ are independent. By standard asymptotic arguments (Serfling, 1980, p. 72), $\mathbf{T}$ is asymptotically bivariate normal with mean $(\sigma^2, \sigma_G^2)^{\mathrm{T}}$ and diagonal covariance matrix diag $\{\mu_4 - \sigma^4, \mu_{4G} - \sigma_G^4\}/n$. Applying the delta method (Serfling, 1980, Corollary 3.3) to $g(\mathbf{T}) = \ln(\mathbf{t}_1) - \ln(\mathbf{t}_2)$ gives the stated result.

Proof of Proposition 3. Let $D_{ij} = X_{ij} - G_i$, and consider the $(J+1)$-dimensional random vector $\mathbf{V}_i = (\mathbf{D}_i^{\mathrm{T}}, G_i)^{\mathrm{T}}$, where $\mathbf{D}_i^{\mathrm{T}} = (D_{i1}, \ldots, D_{iJ})$. (Note that the linear transformation from $(X_{i1}, \ldots, X_{iJ}, G_i)$ to $\mathbf{V}_i$ is invertible.) $\mathbf{V}_i$ is normally distributed with mean vector $\mathrm{E}(\mathbf{V}_i) = (\mathbf{0}_J^{\mathrm{T}}, \mu)^{\mathrm{T}}$, since $\mathrm{cov}(D_{ij}, G_i) = 0$, the covariance matrix of $\mathbf{V}_i$ is block diagonal. The upper $J \times J$ block is $\mathrm{cov}(\mathbf{D}_i) = \boldsymbol{\Sigma}$, and the lower 1×1 block is $\mathrm{var}(G_i) = \sigma_G^2$. Therefore, the random vectors $\mathbf{D}_i$ and G_i are independent.

Thus, the likelihood factors into two components, the first involving only the parameters of $\boldsymbol{\Sigma}$ and depending on $\mathbf{D}_1, \ldots, \mathbf{D}_n$ and the second involving only the parameters μ and σ_G^2 and depending on $G_1, \ldots, G_n$. From this factored likelihood, the maximum likelihood estimators are easily found to be $\hat{\boldsymbol{\Sigma}} = \Sigma \mathbf{D}_i \mathbf{D}_i^{\mathrm{T}}/n, \hat{\mu} = \bar{G}$, and $\hat{\sigma}_G^2 = S_{GG}/n$. In particular, $\hat{\sigma}_{jj} = S_{DD(j)}/n$, where $S_{DD(j)} = \Sigma D_{ij}^2$. By the invariance of maximum likelihood estimators under nonsingular transformations, it follows that $r_{g(j)}^2$ is the maximum likelihood estimator of ρ_j, for $j = 1, \ldots, J$.