Biostatistics (2004), **5**, 3, *pp.* 399–413
doi: 10.1093/biostatistics/kxg043

Comparing and predicting between several methods of measurement

BENDIX CARSTENSEN
Steno Diabetes Center, Niels Steensens Vej 2, Dk-2820 Gentofte, Denmark
bxc@steno.dk

SUMMARY

In studies designed to compare different methods of measurement where more than two methods are compared or replicate measurements by each method are available, standard statistical approaches such as computation of limits of agreement are not directly applicable. A model is presented for comparing several methods of measurement in the situation where replicate measurements by each method are available. Measurements are viewed as classified by method, subject and replicate. Models assuming exchangeable as well as non-exchangeable replicates are considered. A fitting algorithm is presented that allows the estimation of linear relationships between methods as well as relevant variance components. The algorithm only uses methods already implemented in most statistical software.

Keywords: Calibration; Exchangeability; Functional model; Measurement error; Method comparison; Prediction; Ultrastructural model; Variance component model.

1. INTRODUCTION

In epidemiological studies involving several centres, it is customary to encounter clinical measurements made by several different methods, in which case we need to be able to translate measurements between the various methods, and in particular to take account of different sources of error attached to the methods. This will require both conversion formulae as well as estimates of variance components for the measurement methods in question.

Similar needs arise in laboratory studies where a number of measurement methods (or machines) are compared; sources of variation for different methods need to be quantified in order to choose between them, and once a choice has been made the need for accurate conversions between old and new methods are required.

1.1 *A motivating example*

Diabetes patients attending the outpatient clinic at Steno Diabetes Center (SDC) have their HbA_{1c} levels routinely measured at every visit. HbA_{1c} is a marker for the long term glucose-regulation of patients. It is measured as the fraction of haemoglobin being glucosylated—for normal persons the value will be around 5% and the treatment goal for diabetes patients is usually to maintain a value below 7.5%.

In connection with the purchase of a new device for measurement of HbA_{1c} in blood samples at the SDC laboratory, three machines (the existing, `BR.VC`, and two candidates, `BR.V2` and `Tosoh`) were compared. Venous and capillary blood samples were obtained from all patients appearing in the outpatient clinic on two consecutive days who consented to have extra blood samples taken for the experiment. 38

patients gave consent. Samples were measured on four consecutive days on each machine, hence there were five analysis days. All machines were calibrated every day to the manufacturers' standards.

Measurements of HbA_{1c} are thus classified by method (=machine×type of blood), individual (=patient) and replicate (=day of analysis). In this case the replicates are clearly not exchangeable, neither within patients nor simultaneously for all patients.

The aim was to help decide which machine to buy and to produce a reliable prediction between the existing machine and the new one (whichever one was chosen). We also wanted to know about the relationship between measurements made on venous blood samples (from the arm) and capillary blood samples (from the ear lobe).

All pairwise plots of means over days for the six methods (three machines and two types of blood, capillary and venous) and 38 patients are shown in Figure 1.

Our first aim is to produce conversions between methods which, unlike regression analysis, give the same results in either direction. For example, Figure 1 shows that regressing method y =`BR.V2.ven` on x =`Tosoh.ven` gives $y = 0.34 + 0.97x$, whereas the opposite regression gives $y = 0.27 + 0.98x$. These regressions do not use the information from the replicate measurements or the relationships between the other methods for the same persons. Using the methods outlined in this paper we obtain the relationship $y = 0.349 + 0.973x$.

Our second aim is to estimate the components of variation in the measurements by different methods.

1.2 *The Bland–Altman model*

The usual approach to comparing two methods of measurement is the one given by Bland and Altman (1986), where the device of 'Limits of Agreement' is explained and the so-called Bland–Altman plot is introduced.

The Bland and Altman approach assumes that one measurement by each method has been carried out on a number of individuals. The limits of agreement are prediction limits for the difference between measurements by the two methods on a randomly chosen individual.

The model underlying this procedure (restricting attention to normal models) is

$$y_{mi} = \alpha_m + \mu_i + e_{mi}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.1}$$

where y_{mi} denotes a measurement by method m on individual i. This leads to differences, $d_i = y_{1i} - y_{2i}$ being identically distributed with mean $\alpha_1 - \alpha_2$ and variance $\sigma_1^2 + \sigma_2^2$, independent of the averages $\bar{y}_{\cdot i}$ if $\sigma_1 = \sigma_2$. The so-called Bland–Altman plot (d_i versus $\bar{y}_{\cdot i}$) is used to inspect visually whether the difference and its variance is constant as a function of the average.

This model assumes that the only difference between the methods (on the scale chosen), is that one is offset by a constant amount from the other. The model (1.1) is formulated as a two-way analysis of variance model which leads to a paired t-test for equality of the mean method-levels (i.e. testing if the difference is 0). The generalization to several methods of measurement is straightforward. If more than two methods are involved, it is possible to identify the single variance components σ_m.

1.3 *Replicate measurements*

The model (1.1) can also be used if replicate measurements are present, in which case it would be natural to expand it with an extra component of variance, separating the measurement error from the individual×method interaction:

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), \quad e_{mir} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.2}$$

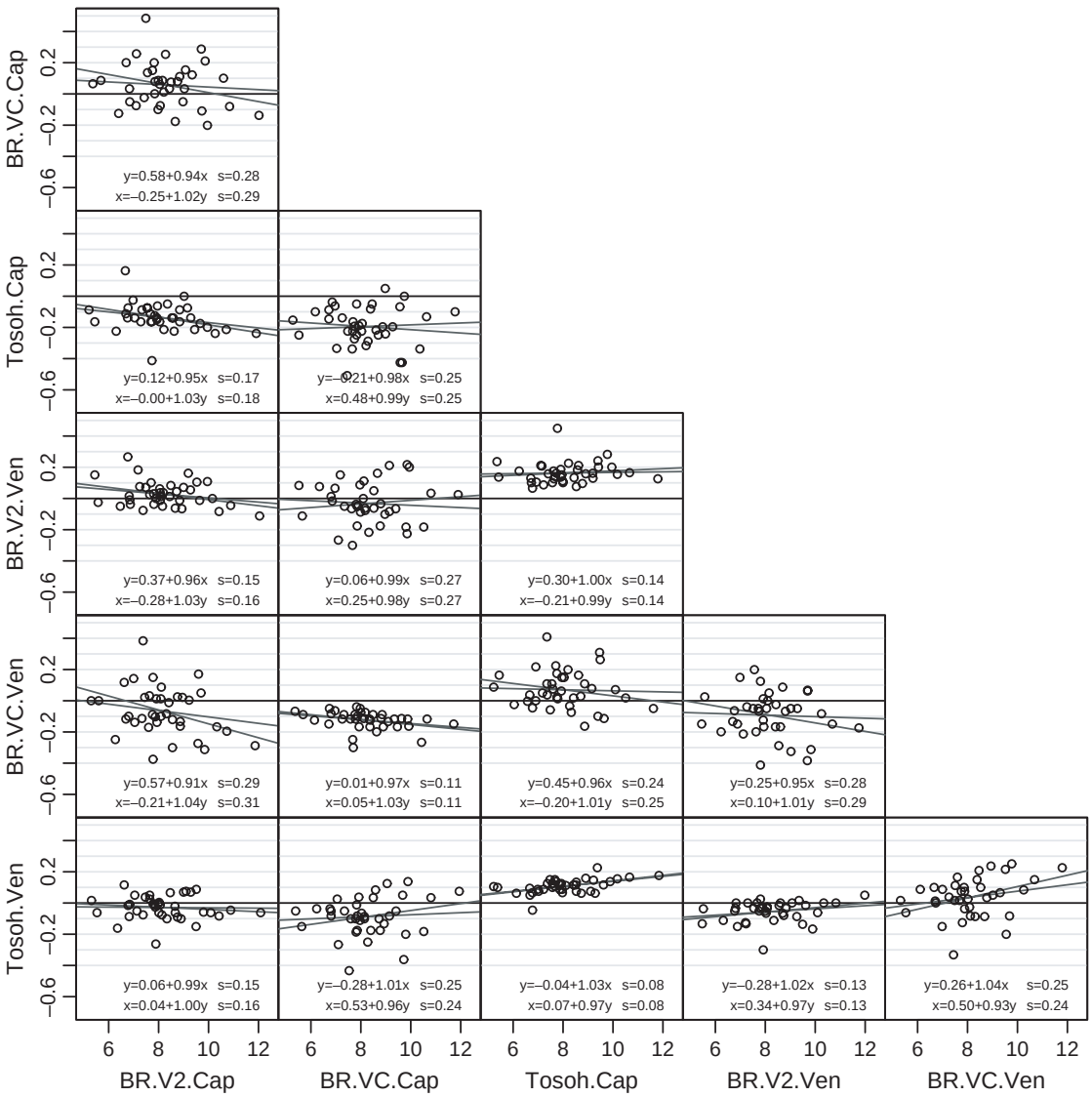

Fig. 1. Averages of HbA_{1c} measurements for the 38 persons over five days by the six different methods considered, compared for all pairs of methods. Each panel is a Bland-Altman plot: $(y-x)/2$ versus $(y+x)/2$—a 45° clockwise rotation of the y vs. x plot. The lines shown are the two regression lines. The formulae for the regression lines and the residual standard deviations are printed in each panel.

with all the random effects assumed independent. In this model it is assumed that replicate measurements are *exchangeable* within each method. If replicates for all methods in a particular individual are done in parallel this assumption does not hold.

The model (1.2) is a two-way analysis of variance model with a random interaction term and separate variances in each column. Separate variances of the interactions (τ_m^2) can only be estimated if at least three measurement methods are compared, whereas separate residual variances (σ_m^2) can always be estimated

if replicate measurements are present. The model can be fitted by standard software packages for mixed models.

1.4 *Extensions*

In the model (1.1) the interaction term is omitted and in (1.2) left unspecified, albeit random, but in measurement method comparison studies parts of this interaction naturally belong in the systematic (fixed) part of the model, for example by allowing methods to have deviations depending linearly on the level of measurement, μ_i. This is most conveniently formulated as

$$y_{mi} = \alpha_m + \beta_m \mu_i + e_{mi}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2).$$

This model is overparametrized as it stands; the μs are only determined up to a linear transformation.

For two measurement methods, this is the classical problem with errors in both variables. Depending on the ratio σ_1/σ_2, the optimal estimate of the line can be anything between the two traditional regression lines. If the ratio of σ_1 and σ_2 is unknown, there is no way out of this, unless we have replicate measurements in the same individual by each method. In that case, it is possible to estimate the variances, and hence the 'correct' regression line.

Another possibility for estimating in this model is to assume some distribution of the μs as, for example, in Dunn and Roberts (1999) which leads to a structural model.

Robust and non-parametric methods are also available, mostly in the case where only two methods are involved, see for example Passing and Bablok (1983, 1984). In the following the attention will be restricted to situations where replicate measurements by each method are available.

In Section 2 a general model is introduced, in Section 3 a practical estimation procedure is outlined, relying mainly on standard statistical methods. Section 4 deals with prediction from one method to another. Sections 5 and 6 discuss possible extensions relaxing some of the assumptions in the general model. In Section 7 a more detailed account of the introductory example is given. The relationship to the ultrastructural model and its variants is discussed in Section 8.

2. A GENERAL MODEL FOR METHOD COMPARISONS

2.1 *Notation and terminology*

Consider the situation where a number of measurement methods are to be compared in order to quantify the precision (sources of variation) for each of them and estimate the relationships between them. An experiment is conducted where for each **item** (blood sample, bacterial isolate, individual, field plot, ...), $i = 1, \dots, I$, and **method**, $m = 1, \dots, M$, a number of **replicate** measurements, $r = 1, \dots, R_{mi}$ is performed.

There is no assumption about the data setup being balanced—it is only assumed that the number of methods and replicates is sufficiently large to make the model identifiable. Observations on measurements by a particular method are assumed exchangeable within item; measurements on the same item are not linked across methods, nor across replicates. Relaxing of these assumptions is discussed later.

2.2 *Model*

Assuming a linear relation among the measurement methods we can set up a model where observations by each method are linked linearly to a common 'true' item value. A model of this type would thus include: fixed effect of each item, $\alpha_m + \beta_m \mu_i$; random item$\times$method effect, $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$; random measurement error, $e_{mir} \sim \mathcal{N}(0, \sigma_m^2)$ and independence between measurement errors.

The variances of the random effects must depend on m, since the different methods do not necessarily measure on the same scale, and different methods naturally must be assumed to have different variances. In studies where different methods actually do measure on the same scale, it will be meaningful to compare the variance components between the methods.

In mathematical terms we have

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir}, \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), \quad e_{mir} \sim \mathcal{N}(0, \sigma_m^2). \tag{2.1}$$

Two crucial assumptions in this model are that replicate measurements are exchangeable within (method, item) and that measurements by different methods are independent given μ_i.

This is a functional model for comparison of measurement methods, similar to the model discussed by Kimura (1992), but without any assumptions about known variance ratios. Dispensing with the assumed knowledge of variance ratios is of course only possible because we assume replicate measurements are available for all methods.

The number of μ_is will in most cases be fairly large compared to the total number of observations, unless there are many replicates or methods. Despite this, in designed method comparison studies it will not generally be reasonable to define the item parameters as random according to some distribution, because items in many cases will be deliberately chosen to span a 'relevant' range of values more or less uniformly.

2.3 *Parameters of the mean*

As the model (2.1) is formulated, not all parameters α_m, β_m, μ_i are identifiable. The μs are only identifiable up to a linear transformation:

$$\mu_i \mapsto a + b\mu_i \quad \Rightarrow \begin{cases} \alpha_m & \mapsto & \alpha_m - \dfrac{\beta_m}{b} a \\ \beta_m & \mapsto & \dfrac{\beta_m}{b}. \end{cases}$$

Since the relation between any two methods of measurement is assumed to be linear, an arbitrary one may be taken to be the reference, with means μ_i, that is transforming the μs using $a = \alpha_{\text{ref}}$, $b = \beta_{\text{ref}}$. It is easily seen that the resulting translation formulae between methods are invariant under linear transformation of the μs.

3. Estimation

For fixed values of μ_i, the model (2.1) is a linear mixed model with separate regressions for each m on μ_i, a random effect of method$\times$item and a residual variance. Since the variances are also specific for each method, the model can be fitted separately for each method.

The best linear unbiased predictor (BLUP) for a specific individual i measured with method m in this model has the form

$$\text{BLUP}_{mir} = \hat{\alpha}_m + \hat{\beta}_m \mu_i + \hat{c}_{mi}. \tag{3.1}$$

The parameters are estimated under the assumption that the μs are the true item values. Since this is not the case we may be able to absorb some of the random method$\times$item interaction into the fixed part by updating the μs.

The expression (3.1) suggests that this can be done by regressing $\text{BLUP}_{mir} - \hat{\alpha}_m = \hat{\beta}_m \mu_i + \hat{c}_{mi}$ on $\hat{\beta}_m$ through the origin with separate slope for each item and weights $\hat{\sigma}_m^{-2}$. The estimated slopes will then be the updated values of the μs.

Thus, estimation could be performed by switching between the two formulations, fixing either set of parameters in turn, alternating between updating $(\alpha, \beta, \tau, \sigma)$ and μ. This procedure will also conveniently circumvent the identifiability problem since the two models fitted are perfectly identifiable. The estimates of the α and β obtained are just an arbitrary set, but the parameters linking methods m and k, say:

$$\alpha_m - \frac{\beta_m}{\beta_k}\alpha_k \qquad \text{and} \qquad \frac{\beta_m}{\beta_k}$$

will be invariant under linear transformation of the μ, and hence the arbitrariness has no influence on these parameters of interest.

3.1 *Practicalities*

The practical implementation of the procedure may proceed as follows:

1. Produce initial estimates of μ_i e.g. as the item-means over all methods and replicates.
2. Fit a mixed model for y_{mir} with μ_i as covariate for each m and a random effect of $m \times i$, and compute BLUPs of the random effects, $\hat{c}_{mi}$. This is the model (2.1) assuming the μ_i are values of known covariates. The variance of the $m \times i$ effect as well as the residual variance should be specific for each method.
3. Update the μ by regressing $\text{BLUP}_{mir} - \hat{\alpha}_m$ on $\hat{\beta}_m$ through the origin with weights $\hat{\sigma}_m^{-2}$.
4. Check for convergence in terms of variance parameters and mean parameters of interest, i.e. $\alpha_m - \alpha_k\beta_m/\beta_k$ and β_m/β_k for some fixed k. If no convergence, go to 2.

The regression on $\hat{\beta}_m$, say, should be understood as a regression on a vector of the same length as the set of observations, with values $\hat{\beta}_m$ for all units with measurements by method m.

If the methods are not on the same scale the algorithm can be started by regressing the item means for each method on the first to obtain initial estimates of α and β and using these estimates to convert all measurements to the same scale, where item means can be formed.

3.2 *Standard errors of parameters*

The standard errors of the regression parameters α_m and β_m produced from the random effects models are *conditional* on the estimated values of the μs, and hence are smaller than those one would obtain by maximizing the likelihood simultaneously over both sets of parameters. The same also applies to the estimates of the μs, but these standard errors are of less interest.

The regression parameters are not of interest by themselves, only in the form $\alpha_m - \alpha_{\text{ref}}\beta_m/\beta_{\text{ref}}$ and $\beta_m/\beta_{\text{ref}}$. The individual sets of α and β are conditionally independent given the μs since they are derived from independent datasets. The dependence comes from the fact that the μs are derived from the total dataset.

An approximate variance of the relative slopes estimated by $\hat{\beta}_m/\hat{\beta}_k$ can be derived by Taylor expansion from the estimated standard errors of the β-estimates, κ_m and κ_k, say:

$$\widehat{\text{var}}\left(\frac{\hat{\beta}_m}{\hat{\beta}_k}\right) \approx \frac{\kappa_m^2}{\hat{\beta}_k^2} + \frac{\kappa_k^2\hat{\beta}_m^2}{\hat{\beta}_k^4} = \frac{1}{\beta_k^2}\left(\kappa_m^2 + \frac{\beta_m^2}{\beta_k^2}\kappa_k^2\right).$$

The first term is what one would get if $\hat{\beta}_k$ is taken as fixed, the second is the correction for the variance of the denominator. This expression is invariant under rescaling of the μs; the same scaling factor will apply both to β and κ. It can be shown that the standard errors of the αs are also invariant under transformation of the μs, in the sense that α_m for a given α_{ref} will have the same standard error regardless of the scaling of the μs prior to the transformation $\alpha_m \mapsto \alpha_m - \alpha_{\text{ref}}\beta_m/\beta_{\text{ref}}$.

4. Prediction

When comparing two methods of measurement with the intent of predicting method 1 from method 2, one may argue that the regression of method 1 on method 2 should be used, since this is based on the *conditional* distribution of y_1 given y_2, which is exactly the prediction situation (Carroll *et al.*, 1995). However, this argument relies heavily on an assumption that the 'new' observation from which the prediction is done is randomly chosen from the same population as was used for the estimation.

In practical situations this will not necessarily be the case, because prediction will typically be needed for populations different from the one used in the calibration study. Otherwise separate calibration studies would be needed for each population.

In most circumstances the calibration sample will (or should) be chosen to give maximal accuracy of the comparison over the range where the conversion is to be used, so the distribution of the variable of interest in the calibration sample is not necessarily close to the population distribution.

Predictions based on the model (2.1), assuming methods to be conditionally independent given μ_i, should include both the measurement variation σ_m and the method×item variation τ_m, but also take the uncertainty in the measurement of the observed value into account.

For a (new) observed value of y_2, y_{20}, say, we have

$$y_{20} = \alpha_2 + \beta_2\mu_0 + c_{20} + e_{20} \qquad \Leftrightarrow \qquad \mu_0 = \frac{y_{20} - \alpha_2 - c_{20} - e_{20}}{\beta_2}$$

which leads to predicting the measurement by method 1, y_{10}, by

$$y_{10} = \alpha_1 + \beta_1\mu_0 + c_{10} + e_{10} = \alpha_1 + \beta_1\frac{y_{20} - \alpha_2 - c_{20} - e_{20}}{\beta_2} + c_{10} + e_{10}.$$

Hence the mean and variance of y_{10} conditional on y_{20} is

$$\mathrm{E}(y_{10}) = \hat{\alpha}_1 + \frac{\hat{\beta}_1}{\hat{\beta}_2}(y_{20} - \hat{\alpha}_2), \qquad \mathrm{V}(y_{10}) = \left(\frac{\hat{\beta}_1}{\hat{\beta}_2}\right)^2(\hat{\tau}_2^2 + \hat{\sigma}_2^2) + (\hat{\tau}_1^2 + \hat{\sigma}_1^2) \tag{4.1}$$

so the prediction variance depends both on the variance on the scale of the predictee as well as on the scale of the predictor. This kind of prediction interval has the property that it will produce a set of prediction bounds in a (y_1, y_2)-plot which is the same regardless of whether y_1 is predicted from y_2 or vice versa—the slope of the line linking y_1 with y_2 is β_1/β_2 so the vertical distance between two lines with this slope is β_1/β_2 times the horizontal, which is exactly the ratio of the standard deviations used in prediction in the two directions.

4.1 *Incorporating the estimation variance*

By analogy with the classical prediction problem from linear regression it is not only the estimated variance that should be used, we should add the variance of the estimated mean, $\hat{\alpha}_1 + (\hat{\beta}_1/\hat{\beta}_2)(y_{20} - \hat{\alpha}_2)$. If $\hat{\Sigma}$ is the 4×4 estimated covariance matrix of $(\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2)$ (conditional on the estimated values of μ_i), then the variance of the prediction mean, $f(\hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2) = \hat{\alpha}_1 + (y_{2i0} - \hat{\alpha}_2)\hat{\beta}_1/\hat{\beta}_2$ is

$$\mathrm{D}f^{\mathrm{T}}\,\hat{\Sigma}\,\mathrm{D}f = \left(1, \frac{(y_{2i0} - \hat{\alpha}_2)}{\hat{\beta}_2}, -\frac{\hat{\beta}_1}{\hat{\beta}_2}, \frac{-\hat{\beta}_1(y_{2i0} - \hat{\alpha}_2)}{\hat{\beta}_2^2}\right)\hat{\Sigma}\begin{pmatrix} 1 \\ (y_{2i0} - \hat{\alpha}_2)/\hat{\beta}_2 \\ -\hat{\beta}_1/\hat{\beta}_2 \\ -\hat{\beta}_1(y_{2i0} - \hat{\alpha}_2)/\hat{\beta}_2^2 \end{pmatrix}.$$

Since methods are assumed independent given true value of μ_i, the matrix Σ will be block-diagonal with 2×2 blocks along the diagonal:

$$\Sigma = \begin{pmatrix} \kappa_{a1}^2 & \rho_1\kappa_{a1}\kappa_{b1} & 0 & 0 \\ \rho_1\kappa_{a1}\kappa_{b1} & \kappa_{b1}^2 & 0 & 0 \\ 0 & 0 & \kappa_{a2}^2 & \rho_2\kappa_{a2}\kappa_{b2} \\ 0 & 0 & \rho_2\kappa_{a2}\kappa_{b2} & \kappa_{b2}^2 \end{pmatrix}.$$

This is not quite true, because the correlation between parameters from different methods induced by the μ_i in the model is ignored, but the approximation is reasonable to see if the uncertainty in (α,β) has any effect on the prediction.

However, if a method comparison study is carefully designed and adequately sized, these corrections will be of minimal importance for the conclusions of the study.

4.2 *Prediction based on replicate measurements*

If there are k replicate measurements of y_2 available, the prediction of y_1 should then be based on the average of these, $\bar{y}_{20\cdot}$. Under the model, the average of the measurements will contain only one value of c_{20}, but k values of e_{mir}, so the variance contribution for $\bar{y}_{20\cdot}$ will be

$$\left(\frac{\beta_1}{\beta_2}\right)^2 \left(\frac{\tau_2^2}{k^2} + \frac{\sigma_2^2}{k}\right).$$

5. RELAXING THE EXCHANGEABILITY ASSUMPTION

The model 2.1 assumes that replicates on the same item are exchangeable within method. This can be a highly unrealistic assumption, e.g. when replicates for technical or logistic reasons are made on separate days or in batches of some kind. In this situation it is reasonable to introduce a random method$\times$replicate effect (i.e. a method by day/batch effect), which leads to the model

$$y_{mir} = \alpha_m + \beta_m\mu_i + c_{mi} + d_{mr} + e_{mir}, \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2), \quad d_{mr} \sim \mathcal{N}(0, \omega_m^2), \quad e_{mir} \sim \mathcal{N}(0, \sigma_m^2). \tag{5.1}$$

The expression for the predictions under the model (5.1) will be the same as those given for the model (2.1), but the expressions for the prediction variance will include extra terms, ω_1 and ω_2, stemming from the random item$\times$replicate interaction.

There is no reason to restrict the method$\times$replicate interaction to be entirely random, one may take part of it as fixed. If for example there is a suspicion that the quantity measured decays by day of analysis, d say, appropriate models could be

$$y_{mir} = \alpha_m + \beta_m(\mu_i + \delta d) + c_{mi} + d_{mr} + e_{mir}$$

or

$$y_{mir} = \alpha_m + \beta_m(\mu_i + \delta_m d) + c_{mi} + d_{mr} + e_{mir},$$

depending on whether the effect was believed to be different between methods or not.

6. RELAXING THE CONDITIONAL INDEPENDENCE ASSUMPTION

The assumption of independence between methods given item may be unrealistic if items represent plots of an experiment and replicates are subsamples from each plot. In this case measurements on the

same item are linked within replicate, so modelling an item×replicate interaction would be appropriate, e.g. by including a random effect for each combination of item and replicate.

But as the model should allow methods to measure on different scales, so the random effect cannot be on the measurement scale, but must be on the μ-scale:

$$y_{mir} = \alpha_m + \beta_m(\mu_i + a_{ir}) + c_{mi} + e_{mir}, \quad a_{ir} \sim \mathcal{N}(0, \nu). \tag{6.1}$$

This model introduces a correlation between observations by different methods on the same item, derived from the linking of replicates across methods (within plots, for example). This correlation is structured by the non-exchangeability of replicates within methods, so that observations by different methods on the same (item, replicate) will be correlated.

An alternative specification for the correlation between measurements by different methods would be to relax the independence assumption for the matrix effects by specifying

$$\text{cov}(c_{mi}, c_{ki}) = \rho_{mk}\tau_m\tau_k \tag{6.2}$$

The structure of this model is most easily compared to the models (5.1) and (6.1) by noting the assumed covariances for the observations:

Model	(2.1)	(5.1)	(6.1)	(2.1) + (6.2)
$\text{cov}(y_{mir}, y_{kir})$	0	0	$\beta_m\beta_k\nu^2$	$\rho_{mk}\tau_m\tau_k$
$\text{cov}(y_{mir}, y_{mjr})$	0	ω_m^2	0	0
$\text{cov}(y_{mir}, y_{mis})$	τ_m^2	τ_m^2	τ_m^2	τ_m^2

The extension in (6.2) is more flexible than that in (6.1), because $M(M-1)/2$ variance parameters are introduced into the model in addition to the $2M$ already there. This is a substantial extension of the number of variance parameters, and will probably require massive amounts of data just to produce reasonably reliable estimates of the ρ_{mk}. It will not be possible to estimate in this model using the algorithm outlined above.

The model (6.1) only introduces one additional parameter, ν^2. If there is a special structure to the replicates it would in principle be possible to extend (6.1) by specifying $\text{var}(a_{ir}) = \nu_r^2$.

For model (6.1) the BLUP-based iterative estimation method does not carry over immediately. Suppose that the μs and the as are known and that the resulting random effects model has been fitted, then

$$\text{BLUP}_{mir} = \hat{\alpha}_m + \hat{\beta}_m(\mu_i + a_{ir}) + \hat{c}_{mi} = \hat{\alpha}_m + \hat{\beta}_m\mu_i + \hat{\beta}_m a_{ir} + \hat{c}_{mi}.$$

A modification of the proposed algorithm would then be to fit a model for $\text{BLUP}_{mir} - \hat{\alpha}_m$ with a fixed term consisting of separate regressions for each i on $\hat{\beta}_m$ (giving estimates of the μ) and a random regression on $\hat{\beta}_m$ with the item×replicate cross-classification as the factor, giving updated values of the a_{ir} as BLUPs from the model

$$z_{mir} = \text{BLUP}_{mir} - \hat{\alpha}_m = \mu_i\hat{\beta}_m + a_{ir}\hat{\beta}_m + e_{mir}$$

fitted with $\hat{\sigma}_m^{-2}$ as weights.

Table 1. *Estimates of variance components for the three different methods. The scale is the standard deviation, i.e. HbA_{1c}%. The sum is the square root of the sum of the squares of the three* stds

	τ (matrix effect)	ω (day to day)	σ (residual)	sum
`BR.V2.Cap`	0.163	0.166	0.092	0.250
`BR.VC.Cap`	0.132	0.030	0.077	0.156
`Tosoh.Cap`	0.113	0.049	0.071	0.142
`BR.V2.Ven`	0.152	0.093	0.075	0.193
`BR.VC.Ven`	0.140	0.020	0.045	0.148
`Tosoh.Ven`	0.107	0.032	0.060	0.127

7. EXAMPLE: MEASUREMENT OF HbA_{1c} AT SDC

We now return to the experiment comparing methods for measurement of HbA_{1c} at SDC. The layout of the experiment involves three machines and two types of specimen, constituting methods $m = 1, \ldots, 6$; individuals $i = 1, \ldots, 38$ and replicate=day $d = 1, \ldots, 5$. The replicates in this experiment are not exchangeable, so it is necessary to include a day×method interaction (separate effect of calibration for each machine):

$$y_{mid} = \alpha_m + \beta_m \mu_i + c_{mi} + d_{md} + e_{mid} \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), \quad d_{md} \sim \mathcal{N}(0, \omega_m^2), \quad e_{mid} \sim \mathcal{N}(0, \sigma_m^2).$$

We might as well have specified the $m \times d$ effect as fixed and estimated the parameters associated with it. We chose to include some of the effect in the fixed part to accommodate systematic changes in measured levels by time (t) since sampling:

$$y_{mid} = \alpha_m + \beta_m(\mu_i + \delta_m t) + c_{mi} + d_{md} + e_{mid}.$$

This has the consequence that the prediction between methods depends on the day of measurement.

From a formal point of view, the variance components cannot be compared in this model, because the measurements y_{mir} are allowed to be on different scales. However since the measurements in this case actually *are* on the same scale (%HbA_{1c}) the comparison makes sense. The variance components are invariant under the linear transformations of the α and β that leaves the model intact, but the values of δ_m are dependent on the chosen parametrization (i.e. scaling of the μ).

The old existing machine at SDC was the `BR.VC`, so the choice was between `BR.V2` and `Tosoh`. From Table 1 it is seen that the `Tosoh` has slightly smaller variance components overall. Of particular interest here is the residual variance representing the repeatability and the method by replicate interaction representing the reproducibility, both of which are seen to be smallest, i.e. best, for `Tosoh`, in particular the latter. Repeatability and reproducibility are further discussed below.

A precise evaluation of whether this is significant or not would require construction of confidence intervals for the variance components. This could be performed by bootstrapping.

The results for the means to be used for future conversions are given as a table with corresponding prediction standard deviations, Table 2 referring to day one after sampling. Entries in this table are used for conversion of clinical measurements, to ensure comparability of measurements within individual patients attending SDC in the transition from the old machine to the new, and to convert between measurements made on venous and capillary blood.

8. Discussion

The models presented here have predecessors that look almost the same, as well as some less related approaches that are mainly designed for use in the case of comparing only two methods with one measurement by each method. The latter are not discussed here, as we only aim at providing modelling tools for use in situations where a comparison experiment with replicates has been conducted.

As stated in equation (2.1), the simplest model is an extension of the functional model (for data without replications) discussed by Kimura (1992). The estimation procedure outlined is also very similar to the EM-algorithm proposed by Kimura, but because of the replicates and the extra variance components the estimation of the μ is not a formal E-step, but an estimation of a subset of the parameters conditional on another subset. Barnett (1970) discusses a structural model for data with replications where the variances are allowed to vary by item, i.e. $\mathrm{var}(e_{mir}) = \sigma_{mi}$, but with the restriction that the ratios σ_{mi}/σ_{ki} were independent of i.

Generally, data from a method comparison study can be arranged in a three-way array classified by method, item and replicate. Since we are dealing with measurements on potentially different scales, any variance component involving method must be allowed different variances across methods.

The pure measurement error will be the three-way interaction in this layout. If replicates are exchangeable it will be the within-cell variation in the two-way array classified by item and method. The three possible two-way interactions can in principle all be estimated, but it is a subject matter decision whether they should enter the model and to what extent they should be included as random or fixed effects.

The method$\times$item interaction is usually split into a parametric part, corresponding to the specification of the linear relationship between methods, and a random method$\times$item effect (matrix effect).

If replicates are non-exchangeable, for example because of simultaneous calibration of machines or batch processing, then a method$\times$replicate interaction should be included, either as random or fixed, or as a mixture of both.

In method comparison studies, where items are physical entities like plots and replicates are subsamples within plots, replicates are made in parallel within items. It may then be necessary to include an item$\times$replicate interaction. Because of the different scales for measurements this interaction cannot be included on the measurement scale, but must be on the μ-scale.

8.1 *Interpretation of variance components and choice between methods*

Throughout it has been assumed that measurements are independent between items given the 'true' value μ_i. The estimates of the μs are essentially weighted means of the measurements by the methods involved. Hence, if some of the methods agree closely because they are subject to the same sources of noise they may dominate the μs.

The variance components are therefore only meaningful under the assumption that the methods compared are measuring the same thing, and that the mean over the methods compared has a sensible interpretation. If for example three methods using one technique for measuring are compared with one method using a different technique, there is a risk of assigning a large variation of the item$\times$method effect to the latter, because of agreement of the item$\times$method effects among the three first. As an extreme example of this, consider a situation where a number of identical methods are compared to one of a different type. The interaction terms for the method$\times$item effects would then be more similar between the three identical machines and would tend to be absorbed in the item parameters (the μs). This would lead to small estimates of the random method$\times$item terms for the three similar methods and larger ones for the last (differing) method, thus giving small variances for the similar methods purely because of the set of methods used.

Table 2. *Table of conversion formulae with prediction standard deviation, based on all persons. These are used to construct prediction intervals: 90% intervals using* $\pm 1.645\times$ std, *95% intervals using* $\pm 1.960\times$ std.
The stds *on the diagonal are* $\sqrt{2\sigma_m^2}$ *for the measurement method, reflecting that predictions are for a repeat sample from the same individual measured by the same method, hence excluding the person×method and day×method variations*

	Predictor, x					
Predicted, y	`BR.V2.Cap`	`BR.VC.Cap`	`Tosoh.Cap`	`BR.V2.Ven`	`BR.VC.Ven`	`Tosoh.Ven`
`BR.V2.Cap`	$0.000 + 1.000x$	$-0.476 + 1.038x$	$-0.011 + 1.041x$	$-0.225 + 1.031x$	$-0.283 + 1.064x$	$0.135 + 1.003x$
	(0.131)	(0.303)	(0.297)	(0.322)	(0.305)	(0.281)
`BR.VC.Cap`	$0.458 + 0.964x$	$0.000 + 1.000x$	$0.448 + 1.003x$	$0.242 + 0.993x$	$0.186 + 1.026x$	$0.588 + 0.967x$
	(0.292)	(0.108)	(0.211)	(0.247)	(0.218)	(0.197)
`Tosoh.Cap`	$0.011 + 0.960x$	$-0.446 + 0.997x$	$0.000 + 1.000x$	$-0.205 + 0.990x$	$-0.261 + 1.022x$	$0.140 + 0.963x$
	(0.285)	(0.210)	(0.100)	(0.239)	(0.207)	(0.186)
`BR.V2.Ven`	$0.218 + 0.970x$	$-0.243 + 1.007x$	$0.207 + 1.010x$	$0.000 + 1.000x$	$-0.056 + 1.033x$	$0.349 + 0.973x$
	(0.313)	(0.249)	(0.241)	(0.106)	(0.248)	(0.227)
`BR.VC.Ven`	$0.266 + 0.939x$	$-0.181 + 0.975x$	$0.255 + 0.978x$	$0.054 + 0.968x$	$0.000 + 1.000x$	$0.392 + 0.942x$
	(0.286)	(0.212)	(0.203)	(0.241)	(0.063)	(0.188)
`Tosoh.Ven`	$-0.134 + 0.997x$	$-0.609 + 1.035x$	$-0.145 + 1.038x$	$-0.359 + 1.027x$	$-0.416 + 1.061x$	$0.000 + 1.000x$
	(0.280)	(0.203)	(0.194)	(0.233)	(0.200)	(0.085)

This is essentially an unidentifiability feature of the method×item interaction. The 'true' matrix effects are not estimable in a design such as the one outlined, but will be confounded with the item parameters (the μs).

If a method×replicate effect is included in the model, similar problems may appear if e.g. daily calibrations are more similar between some methods than others.

The estimates of residual variation are obtainable because they are based on variation between replicates within (item, method).

If replicates are not exchangeable, the method×replicate interaction will represent the reproducibility (ISO 5725-1, 1994), that is the variation between measurements on the same item by the same method under different circumstances (typically, different laboratories). The residual variation will be the repeatability, that is the variation between measurements made on the same item under similar conditions (same method, machine, laboratory, technician). It should be noted that the definitions of repeatability and reproducibility are to some extent subject-matter related, for example with respect to what one would deem to be the 'same equipment within short intervals of time' (ISO 5725-1, 1994).

If replicates are exchangeable, the reproducibility is not available since observations on the same item with the same method under different conditions are not made. The repeatability will be either the residual variation or the sum of the residual variation and the method×replicate variation, depending on the nature of the replications.

If more precise estimates of repeatability and reproducibility were required, one might consider a more complex replication scheme, with some replicates under identical and some under differing conditions. The modelling in this kind of design would in principle be possible along the same lines as described here, for example by incorporating systematic effects in the description of replicates.

The systematic part of the method×item interaction is the linear relationship between the methods, which also link the scales of the methods ($y_{mi} = \alpha_m + \beta_m \mu_i$). Hence, in the comparison of variance components between methods it is necessary to rescale in order to make the comparisons meaningful, e.g. by using terms such as τ_m/β_m, σ_m/β_m etc. In many practical settings where methods measuring on the same scale are compared, all the β will be close to 1 (or more precisely, will be similar) so the rescaling of variance components will have little effect on the comparisons.

In summary, the residual variance and the method×replicate interactions will be the variance components of major interest since they represent the repeatability and reproducibility of the methods (depending on the replication scheme), whereas the method×item interaction, the matrix effects, should be used with some care in the comparison of methods because the relative sizes of these between methods may be influenced by the set of methods compared.

8.2 *The ultrastructural model and variants*

The ultrastructural model was proposed by Dolby (1976) for repeated *pairs* of measurements with two different methods on item i:

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix}, i = 1, \ldots, I,$$

and possibly replicates for each item.

In this model, observations by the same method of measurement on the same item are *not* exchangeable. But the non-exchangeability is due to simultaneous measurement by all methods for each item, which is modelled by a random item×replicate effect, a_{ir}:

$$\begin{aligned} x_{ir} &= \mu_i + a_{ir} + e_{ir1} \\ y_{ir} &= \alpha + \beta(\mu_i + a_{ir}) + e_{ir2} \end{aligned} \qquad a_{ir} \sim \mathcal{N}(0, \phi), \quad e_{irm} \sim \mathcal{N}(0, \sigma_m^2) \tag{8.1}$$

leading to the joint distribution:

$$\begin{pmatrix} x_{ri} \\ y_{ri} \end{pmatrix} \sim \mathcal{N}\left\{\begin{pmatrix} \mu_i \\ \alpha + \beta\mu_i \end{pmatrix}, \begin{pmatrix} \phi + \sigma_1^2 & \beta\phi \\ \beta\phi & \phi + \sigma_2^2 \end{pmatrix}\right\}$$

i.e. with correlation within pairs of measurements for the *same* i.

This model assumes correlation between measurements from different methods within item derived from random effect for both measurement methods, i.e. a random effect associated with the replication of the *pair* of measurements. Thus, in the ultrastructural model replicates are not linked between items within method, but between methods within item, as opposed to the model (2.1).

Rephrasing the ultrastructural model in the spirit of (2.1) by putting $x_{ir} = y_{1ir}$, $y_{ir} = y_{2ir}$, $\alpha_1 = 0$ and $\beta_1 = 1$ we have

$$y_{mir} = \alpha_m + \beta_m(\mu_i + a_{ir}) + e_{mir} \qquad a_{ir} \sim \mathcal{N}(0, \phi), \quad e_{irm} \sim \mathcal{N}(0, \sigma_m^2) \tag{8.2}$$

which is the model (6.1) without the item×method interaction. Here the random effect of replicate is on the μ-scale, i.e. its effect is proportional to β.

The ultrastructural model of Dolby is discussed in a practical setting by Skovgaard (1995) where two methods of measuring flavours in beef are compared over a number of different storage times, t (corresponding to items in the notation in this paper), with replicates being different packs of beef. The random item×replicate is thus a random time×pack effect. The reason for choosing this structure of the model is not quite clear, since the replicates are not subsamples in the sense outlined as example in Dolby's paper. The argument seems to be that packs may age differently. Skovgaard estimates the slope in the ultrastructural model by the ratio of the canonical correlations from a model with storage time as a categorical covariate, i.e. explicitly estimating μ_t in a one-way ANOVA and then using the residuals for estimation of the variance.

The structural model is discussed by Dunn and Roberts (1999) This is a simplification of the ultrastructural model where the μ_i are assumed to come from a normal distribution with a common mean:

$$y_{mi} = \alpha_m + \beta_m\xi_i + e_{mi}, \qquad \xi_i \sim \mathcal{N}(\mu, \phi), \quad e_{mi} \sim \mathcal{N}(0, \sigma^2).$$

The distributional assumption for the ξ induces the correlation between measurements by different methods on the same item. Note that the estimation procedures proposed for this model involves the population variance of the measurements (the variation of the item means, ϕ). This seems strange, because a method comparison study ideally should produce results independent of the population of items used.

8.3 *Summary*

Models for method comparison studies should refer to the data layout and in terms of the subject matter address carefully (1) what effects should be included in the model and (2) whether they should be included as random or fixed. In particular it is not advisable to focus on problems of identifiability of parameters, but rather on model structure.

In reporting results from method comparison studies is important to put the results in an immediately applicable form as conversion tables or charts between methods with proper prediction standard deviations. Further, the size of the variance components should be reported in a form that makes it possible to use the information for comparison of precision between methods.

REFERENCES

BARNETT, V. D. (1970). Fitting straight lines—the linear functional relationship with replicated observations. *Applied Statistics* **19**, 135–144.

BLAND, J. M. AND ALTMAN, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310.

CARROLL, R. J., RUPPERT, D. AND STEPHANSKY, L. A. (1995). *Measurement Error in Non-linear Models*. London: Chapman and Hall.

DOLBY, G. (1976). The ultrastructural relation: a synthesis of the functional and structural relations. *Biometrika* **63**, 39–50.

DUNN, G. AND ROBERTS, C. (1999). Modelling method comparison data. *Statistical Methods in Medical Research* **8**, 161–179.

ISO 5725-1 (1994). Accuracy (trueness and precision) of measurement methods and results—Part 1: General principles and definitions. *International Organization for Standardization* http://www.iso.ch.

KIMURA, D. K. (1992). Functional comparative calibration using an EM algorithm. *Biometrics* **48**, 1263–1271.

PASSING, H. AND BABLOK, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison in clinical chemistry, part I. *Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 709–720.

PASSING, H. AND BABLOK, W. (1984). Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison in clinical chemistry, part II. *Journal of Clinical Chemistry and Clinical Biochemistry* **22**, 431–445.

SKOVGAARD, I. M. (1995). Modelling relations between instrumental and sensory measurements in factorial experiments. *Food Quality and Preference* **6**, 239–244.

[*Received November 18, 2002; first revision April 25, 2003; second revision October 22, 2003; third revision November 19, 2003; accepted for publication December 17, 2003*]