The analysis of very small samples of repeated measurements
II: A modified Box correction

Simon S. Skene[a*†] and Michael G. Kenward[b]

There is a need for appropriate methods for the analysis of very small samples of continuous repeated measurements. A key feature of such analyses is the role played by the covariance matrix of the repeated observations. When subjects are few it can be difficult to assess the fit of parsimonious structures for this matrix, while the use of an unstructured form may lead to a serious lack of power. The Kenward-Roger adjustment is now widely adopted as a means of providing an appropriate inferences in small samples, but does not perform adequately in very small samples. Adjusted tests based on the empirical sandwich estimator can be constructed that have good nominal properties, but are seriously underpowered. Further, when such data are incomplete, or unbalanced, or non-saturated mean models are used, exact distributional results do not exist that justify analyses with any sample size. In this paper, a modification of Box's correction applied to a linear model based F-statistic is developed for such small sample settings and is shown to have both the required nominal properties and acceptable power across a range of settings for repeated measurements.

Keywords: ANOVA; Box correction; covariance matrix; linear model; repeated measures; Scheffés method; small samples

1. Introduction

The preceding companion paper to this [1] highlights the need for appropriate methods for the analysis of repeated measurements when the sample size is very small. In particular, deficiencies in conventional approaches to inference from such samples are highlighted where data are unbalanced or incomplete. It is shown that hypothesis tests based on conventional Wald type procedures do not approximate their nominal properties sufficiently well. This is the case with the Kenward-Roger adjustment [2, 3], which is implemented in *SAS PROC MIXED* [4] and accounts for both bias and variability in the estimated covariance matrix of the fixed effects by adoption of a scaled F-statistic and an adjustment to the denominator degrees of freedom. Also considered is the empirical sandwich estimator from the generalized estimating equations (GEE) approach for categorical data, which uses ordinary least squares

[a] *Department of Economics, Mathematics and Statistics, Birkbeck, London, UK.*

[b] *Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.*

* *Correspondence to: Simon S. Skene, Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.*

† *E-mail: s.skene@bbk.ac.uk*

estimates for the fixed effects and adjusts their standard errors to reflect the observed dependencies in the data. Such a procedure is known to have poor small sample properties, but an adjusted Wald test due to Pan and Wall [5] can be generalised to allow testing of any general linear hypothesis involving fixed effects. When combined with a bias adjustment (Mancl and DeRouen [6]), the resulting statistic is seen to achieve adequate control over the type 1 error rate, but has very poor power in comparison to the Kenward-Roger adjusted statistic (when the latter provides a valid comparator). In this paper we consider further the need for an appropriate general approach, which will be applicable across a range of settings for repeated measures data where the sample size is very small.

The problems in the methods listed above all stem from the same source: the lack of information in the data on the covariance structure of the repeated measurements. In certain balanced settings with saturated mean models the covariance structure does not influence the parameter estimates which are then identical to those obtained from ordinary least squares. In such circumstances the Kenward-Roger adjustment recovers exact Hotelling's T^2 and hierarchical ANOVA F-tests which do not require small sample approximation. Moving away from such special cases, such as with missing data or with non-saturated models, leads to serious departures from the nominal properties of inference procedures in very small samples. It is clear from this that removing the estimated covariance structure from the estimation of the regression parameters leads to an improvement in the small sample behaviour of inferences. In such cases, the estimated covariance structure is not used in the estimation of the mean parameters, but is necessary for estimates of their precision. This control is seen to deteriorate where the covariance structure enters the estimation step, and is worse still in situations which are unbalanced. This suggests the use of ordinary least squares more widely in such settings, with subsequent correction of the measure of precision using a sandwich (also termed empirical or robust) estimate of the covariance matrix of the estimates. In the companion paper it was shown how such procedures, with appropriate small sample modification of the sandwich estimator, produce test statistics with good approximation to their nominal size. However, the need to estimate the covariance matrix of the repeated measurements as part of the sandwich estimator leads to very poor precision of this, with a consequent impact on the power of the associated tests. This was seen to be unacceptably poor in small sample settings.

To overcome this, a natural next step is to remove the estimated covariance structure from both the estimates and their estimates of precision (which will then be strictly invalid). The estimated covariance matrix is then introduced only at the final stage to approximate the sampling distribution of the resultant independence-based statistics, and which forms the basis of subsequent inferences. This idea is not new and forms the basis of the well-known Box correction [7, 8], which suggests a modification to the one-way ANOVA F-statistic to account for departures from this model. Its use in the context of cross-over designs was suggested by Bellavance *et al* [9] who show that Box's modified F-test approximation gives adequate control over the type 1 error rate in this context (see also [10]), but it is considered here in a general linear model setting. We find that a modification to Box's correction is needed to provide acceptably good behaviour across a range of very small sample repeated measurements settings. We introduce Box's correction in Section 2, and

develop the modification in Section 3. The properties of the corrections are explored in Section 4. In Section 5 we compare, through a series of simulation studies across a range of repeated measurement settings, the performance of the corrections with the procedures based on the Kenward-Roger adjustment and the adjusted sandwich estimator. Examples of the use of the modified Box correction are presented in Section 6, where it is also shown how Scheffé's method [11] for obtaining confidence intervals for individual contrasts can be adapted to allow for the modified Box-corrected statistic. Practical recommendations for the analysis of very small samples of repeated measurements are then made in the discussion of Section 7.

2. Box's correction

We suppose that the data can be represented by a conventional multivariate Gaussian linear model [12]. Let $\mathbf{y}_i$ $(T_i \times 1)$ be the response vector from the ith of n subjects, and it is assumed that a common set of measurement times applies to all subjects, although not all may be observed for all subjects. The model has the following general form

$$\mathbf{y}_i \sim \mathrm{N}\left(\mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_i\right), \quad i = 1, \ldots, n, \tag{1}$$

for $\boldsymbol{\beta}$, $(p \times 1)$, the vector of fixed effect parameters, $\mathbf{X}_i$, $(T_i \times p)$, the design/covariate matrix, and $\boldsymbol{\Sigma}_i$ the $(T_i \times T_i)$ covariance matrix for this subject. Depending on the setting, the covariance matrix can in principle take many forms, including those induced by a random effects structure. Defining $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$, setting $\boldsymbol{\Sigma} = \mathrm{block\text{-}diagonal}\{\boldsymbol{\Sigma}_i\}$ we have the equivalent expression for the whole data set:

$$\boldsymbol{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}; \boldsymbol{\Sigma}\right). \tag{2}$$

Now suppose that we wish to test the null hypothesis that c elements of $\boldsymbol{\beta}$ are zero, but not under the model assumptions above, but rather assuming (wrongly) independent equally variable Gaussian errors. Define $\mathbf{X}_R$ $(n \times (r-c))$ to be the design matrix for the model with the terms to be tested removed, and set $\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{B} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - \mathbf{X}_R(\mathbf{X}_R^T\mathbf{X}_R^T)^{-1}\mathbf{X}_R^T$. Then, using the extra sums of squares principle, the appropriate test statistic for the null hypothesis under the independence assumptions is given by

$$F = \frac{n-r}{c}\frac{\mathbf{y}^T\mathbf{B}\mathbf{y}}{\mathbf{y}^T\mathbf{A}\mathbf{y}}, \tag{3}$$

which, if these assumptions held, would have a null $\mathrm{F}_{c,n-r}$ distribution. Under the more general model (2), Box [7, 8] showed that $\psi^{-1}F$ has an approximate $\mathrm{F}_{v1,v2}$ distribution, where the degrees of freedom and ψ are obtained as follows. The key approximation [13] is to treat the numerator and denominator quadratic forms as having independent scaled chi-squared distributions of the form,

$$Q = \mathbf{y}^T\mathbf{C}\mathbf{y} \overset{\mathrm{approx}}{\sim} g\chi_h^2, \tag{4}$$

where the constant and degrees of freedom parameters g and h are chosen by matching the first and second moments. That is

$$gh = \text{tr}(\mathbf{C}\boldsymbol{\Sigma})$$

$$2g^2h = 2\text{tr}(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\Sigma})$$

From this we get

$$\psi = \frac{(n-r)}{c}\frac{\text{tr}(\mathbf{B}\boldsymbol{\Sigma})}{\text{tr}(\mathbf{A}\boldsymbol{\Sigma})} \tag{5}$$

$$v_1 = \frac{\{\text{tr}(\mathbf{B}\boldsymbol{\Sigma})\}^2}{\text{tr}\{(\mathbf{B}\boldsymbol{\Sigma})^2\}} \tag{6}$$

$$v_2 = \frac{\{\text{tr}(\mathbf{A}\boldsymbol{\Sigma})\}^2}{\text{tr}\{(\mathbf{A}\boldsymbol{\Sigma})^2\}} \tag{7}$$

In practice, any consistent estimator of $\boldsymbol{\Sigma} = \text{V}(\boldsymbol{y})$, may be used to compute the adjusted F-distribution parameters. Jones and Keward [14] suggest the use of the ordinary least squares (OLS) covariance estimate is in keeping with the spirit of this approach. However, for data which are unbalanced or have missing values, so that the OLS and restricted maximum likelihood (REML) estimates do not coincide, it may be more practical to simply adopt the unstructured REML estimate, which is widely implemented in existing software. In cases where an unstructured REML estimate cannot be computed (as can occur, for example, where there are too many measurements on too few subjects), $\boldsymbol{\Sigma}$ may be taken to be the most complex covariance structure that the data will support, such as a (high order) antedependence structure. An advantage of this method is that it does not require a non-singular estimate of the covariance structure, so that, in such cases, it is possible to proceed using an 'empirical' estimator such as the sample covariance matrix.

In simulations (see below) Box's correction can be shown to be conservative, giving excessive control over the type 1 error rate and resulting in test sizes well below the nominal rate, and so a modification to this correction is proposed.

3. A modified Box correction

Rather than approximating the distribution of the quadratic form in the F-statistic (3)

$$\frac{Q_1}{Q_2} = \frac{\mathbf{y}^T\mathbf{B}\mathbf{y}}{\mathbf{y}^T\mathbf{A}\mathbf{y}}$$

as a ratio of independent scaled chi-squared distributions, we instead approximate it directly by a scaled F-distribution, $\lambda F_{v_1,v_2}$, and match the first two moments of this. These can be obtained using the the 'delta' method (see, for example, Stuart and Ord [15]), from which we obtain

$$\text{E}\left(\frac{Q_1}{Q_2}\right) \approx \frac{\text{E}(Q_1)}{\text{E}(Q_2)} \tag{8}$$

4

and

$$\text{Var}\left(\frac{Q_1}{Q_2}\right) = \frac{\{\text{E}(Q_1)\}^2}{\{\text{E}(Q_2)\}^2}\left[\frac{\text{Var}(Q_1)}{\{\text{E}(Q_1)\}^2} + \frac{\text{Var}(Q_2)}{\{\text{E}(Q_2)\}^2} - \frac{2\text{Cov}(Q_1, Q_2)}{\text{E}(Q_1)\text{E}(Q_2)}\right] \tag{9}$$

Assuming, as in the Box correction, that the numerator and denominator terms in the F-statistic are independent, so that $\text{Cov}(Q_1, Q_2) = 0$, we have, equating these moments with those of the scaled F-distribution

$$\frac{1}{\lambda}\frac{\text{tr}(\mathbf{B\Sigma})}{\text{tr}(\mathbf{A\Sigma})} = \frac{v_2}{v_2 - 2} \tag{10}$$

and,

$$\frac{1}{\lambda^2}\frac{\{\text{tr}(\mathbf{B\Sigma})\}^2}{\{\text{tr}(\mathbf{A\Sigma})\}^2}\left[2\frac{\text{tr}\{(\mathbf{B\Sigma})^2\}}{\{\text{tr}(\mathbf{B\Sigma})\}^2} + 2\frac{\text{tr}\{(\mathbf{A\Sigma})^2\}}{\{\text{tr}(\mathbf{A\Sigma})\}^2}\right] = \frac{2v_2^2(v_2 + v_1 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \tag{11}$$

Fixing $v_1 = c$, the dimension of the test (similarly to the Kenward-Roger and small-sample sandwich adjustments), we can use these final two equations to obtain expressions for the scale factor λ and the denominator degrees of freedom v_2 for our approximating distribution. This gives us

$$\text{F} = \frac{(n-r)}{c}\frac{\mathbf{y}^T\mathbf{By}}{\mathbf{y}^T\mathbf{Ay}} \overset{\text{approx}}{\sim} \lambda\text{F}_{c,v_2} \tag{12}$$

where,

$$\lambda = \frac{(n-r)}{c}\frac{v_2 - 2}{v_2}\frac{\text{tr}(\mathbf{B\Sigma})}{\text{tr}(\mathbf{A\Sigma})} \tag{13}$$

$$v_2 = \frac{c(4\text{V} + 1) - 2}{c\text{V} - 1} \tag{14}$$

and,

$$\text{V} = \left[\frac{\text{tr}\{(\mathbf{B\Sigma})^2\}}{\{\text{tr}(\mathbf{B\Sigma})\}^2} + \frac{\text{tr}\{(\mathbf{A\Sigma})^2\}}{\{\text{tr}(\mathbf{A\Sigma})\}^2}\right] \tag{15}$$

Results from simulations showing the performance of this statistic across a range of settings for repeated measurements will be shown in Section 5. A further modification was considered using second order deviations about the mean in (8). That is, taking

$$\text{E}\left(\frac{\mathbf{y}^T\mathbf{By}}{\mathbf{y}^T\mathbf{Ay}}\right) \approx \frac{\text{tr}(\mathbf{B\Sigma})}{\text{tr}(\mathbf{A\Sigma})}\left[1 + 2\frac{\text{tr}\{(\mathbf{A\Sigma})^2\}}{\{\text{tr}(\mathbf{A\Sigma})\}^2} - 2\frac{\text{tr}(\mathbf{A\Sigma B\Sigma})}{\text{tr}(\mathbf{A\Sigma})\text{tr}(\mathbf{B\Sigma})}\right] \tag{16}$$

and proceeding as above (with $\text{tr}(\mathbf{A\Sigma B\Sigma}) = 0$). However, simulations show that there is little advantage in following this more complicated approach, since this results in an adjusted statistic which is more conservative, with test sizes closer to those given by the Box correction which we are attempting to inflate.

4. Properties of the Box corrections

It is worth noting that under the assumption of independence, Box's original correction recovers the ANOVA F-statistic exactly, whereas the modified correction does not, although the disparity is of small order. Although it is desirable to have a statistic which recovers the exact test in appropriate circumstances, such circumstances are unlikely to arise in practice in the context of repeated measurements.

It is also useful to consider how such corrections behave under the assumption of 'compound symmetry', and the relationship between this approach and repeated measures ANOVA. The latter approach was often adopted for 'practical' analyses before modern computing power allowed widespread access to the multivariate general linear model, which is now more commonly used for such data. See, for example, [16] for an outline to this approach, and [17] for more detail.

The general formulation is to treat 'time' (occasions of measurement) as an additional within-subjects factor, and to model the jth measurement on the ith subject as

$$y_{ij} = \mu_{ij} + b_i + e_{ij} \tag{17}$$

where μ_{ij} are suitably specified fixed effects, $b_i \sim N(0, \sigma_b^2)$ are random (individual specific) subject effects, and $e_{ij} \sim N(0, \sigma^2)$ are the usual error terms.

Two sources of variation, between-subjects and within-subjects, lead to a compound symmetry covariance structure for the repeated measurements, with a constant variance, $\sigma_b^2 + \sigma^2$, on the diagonal and a constant covariance off the diagonal, and hence a constant correlation between any pair of repeated measurements on the same subject, given by the intra-class correlation

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \tag{18}$$

Such an approach is appropriate when supported by randomization, such as in a randomised block, or split-plot, design, where the subjects are considered as blocks or main-plots respectively, and in extended examples involving more error strata, [18, 19]. However, this approach is not appropriate for the majority of repeated measurements studies because time cannot be randomized. Also, departures from constant correlation and constant variance are often, although not invariably, observed [12].

Box [8] suggested that departures from compound symmetry could be accounted for by reducing the degree of freedom parameters for the two-way ANOVA F-ratio for the 'occasions'/'time' effect by a multiplicative factor ϵ. The parameter ϵ, $1/(p-1) \le \epsilon \le 1$, is estimated using the sample covariance matrix. This approach was extended by Greenhouse and Geisser to the split-plot design [20] and to the (multivariate) setting of profile analysis [21], through an adjustment to the MANOVA test statistic. Huynh and Feldt [22] show that

in general for such a model to be appropriate it is necessary only for the within-subjects covariance matrix to comply with the assumption of sphericity. This is a less restrictive condition than compound symmetry, since sphericity requires only that the variance of differences in a within-subjects design are equal across all groups. The Greenhouse and Geisser correction uses

$$F \overset{\text{approx}}{\sim} F_{\epsilon(p-1), \epsilon(m-g)(p-1)} \tag{19}$$

where p is the number of repeated measurements on a subject, m is the number of subjects, and g is the number of treatment groups. Adjusted tests using ϵ defined by both Greenhouse and Geisser and Huynh and Feldt are widely implemented in software packages.

Using the Box correction of Section 2, under the assumption of compound symmetry, we find

$$\psi = \frac{\sigma^2}{\sigma_b^2 + \sigma^2} = 1 - \rho \tag{20}$$

That is, the Box correction adjusts the one-way ANOVA F-statistic for 'treatments' to that which would be obtained from the (more restrictive) two-way setting, using 'time' as the other factor. This is equivalent to using an (unadjusted) Wald statistic for the regression parameters in a multivariate linear model with a compound symmetry covariance structure. However, although the numerator degrees of freedom are fixed, the denominator degrees of freedom are lower than from their two-way counterpart, since we are accounting for departures from independence. A similar relationship is found with the modified Box correction under compound symmetry.

In the context of very small samples of repeated measurements, the one-way ANOVA approach with a suitable correction is preferred, since it is more widely applicable across a range of settings. The Greenhouse and Geisser approach, based on the split-plot design, is simply too restrictive to be of use generally, since it requires complete and balanced data and would not, for example, accommodate missing data or cross-over designs.

5. Simulation studies

To investigate the properties of the Box corrections it is appropriate to consider their use over a range of settings for repeated measurements. Consideration is given to the following study designs based on the simple repeated measurement and cross-over designs used to assess the existing repeated measurements methods in [1]. These simulations accommodate both a range of numbers of time points (T) and subjects (n), and include data arising from a wider variety of plausible non-stationary covariance structures. The extended simulation designs are detailed below. For the simple repeated measures designs, we have

(A′) A simple repeated measures experiment, with n subjects randomly allocated to two treatment groups (of equal size), and a response recorded for each subject at each of T time points.

Table I. **A cross-over design for nine treatments (A, B, C, D, E, F, G, H, I).**

Subject	Period								
	1	2	3	4	5	6	7	8	9
1	A	B	C	D	E	F	G	H	I
2	B	D	A	F	C	I	H	G	E
3	C	F	E	G	D	B	I	A	H
4	D	G	F	I	B	H	E	C	A
5	E	A	I	C	H	D	F	B	G
6	F	H	B	E	I	G	A	D	C
7	G	I	D	H	F	A	C	E	B
8	H	C	G	B	A	E	D	I	F
9	I	E	H	A	G	C	B	F	D
10	I	H	G	F	E	D	C	B	A
11	E	G	H	I	C	F	A	D	B
12	H	A	I	B	D	G	E	F	C
13	A	C	E	H	B	I	F	G	D
14	G	B	F	D	H	C	I	A	E
15	C	D	A	G	I	E	B	H	F
16	B	E	C	A	F	H	D	I	G
17	F	I	D	E	A	B	G	C	H
18	D	F	B	C	G	A	H	E	I

(B′) As design (A′), but with missing values. An equal number of subjects in each treatment group drop out at some random time following the first observation.

(C′) A five treatment-five period cross-over trial, with $n = 10$ and 20 subjects allocated randomly to treatments according to Table I of [1], using a pair of Williams' squares.

(D) A nine treatment-nine period cross-over trial, with $n = 18$ and 36 subjects allocated randomly to treatments according to Table I of this paper.

In designs (A′) and (B′), we consider $T = 5$ time points with $n = 10$ and 20 subjects, and $T = 10$ time points with $n = 20$ and 40 subjects. Additionally, in design (B′), the numbers of subjects allowed to drop out are given in Table II, below. For the extended simulations involving additional subjects in the cross-over designs ($n = 20$ in design (C′), and $n = 36$ in design (D)), the allocation tables are simply repeated.

Table II. **Number of drop out subjects in extended study design (B′).**

Number of measurements per subject (T)	Number of subjects (n)	Number of subjects to drop out
5	10	2
	20	4
10	20	4
	40	8

Data samples are independently generated arising from a Gaussian distribution with zero mean for a number of underlying covariance structures, and the appropriate null hypotheses under consideration are those of no treatment/time interaction in designs (A′) and (B′),

and no treatment effect in designs (C) and (D). Two underlying stationary covariance structures, compound symmetry and AR1 (high correlation) are considered, together with three non-stationary structures, heterogeneous compound symmetry, heterogeneous AR1 and first order independence. For each of the non-stationary structures, variances are restricted so that they differ by no more than a factor of 10 over the range of the measurement times. These structures are shown in the Appendix.

Results from the extended simulations of study designs (A′) and (B′) are shown in Tables III and IV respectively. The results of simulations concerning study designs (C′) and (D) are shown in Table V.

For $T = 5$ time points in design (A′), we see that as the number of subjects rises from 10 to 20, the power of the tests using the Box corrections is above that of the KR adjusted Wald test for data arising from the stationary covariance structures. However, for data arising from the non-stationary structures, the increase in power is generally lower in comparison to the level attained by the KR adjusted test. This is particularly noticeable for the data arising from the antedependence structure, which is furthest from the univariate linear model assumptions of independence and homogeneity of variance. This is as we might expect, the performance of the KR adjusted test improving as the sample size increases, and the Box corrections performing comparatively less well for large departures from independence. A similar pattern is observed in design (B'), although the loss of power relative to the KR adjusted Wald test is more apparent for increased sample sizes where we have missing values.

As the number of time points increases to $T = 10$, the loss in power of the Box corrections relative to the KR adjusted test is less marked. That is, as the number of subjects and time points is increased in the balanced and complete data setting of design (A′), the modified Box correction appears to hold its own against the KR adjusted Wald test. For design (B′), where the missing values introduce imbalance, the KR adjustments no longer give an exact Hotelling T^2 test, and must be calculated individually for each data set. For the (10×10) matrices necessitated by considering $T = 10$ time points, this is too expensive in terms of available computational time for such a practical study. In order to provide a comparison, the KR adjusted test results in Table IV have been estimated in each case using the 'known' underlying covariance structure and an average number of observations. This gives a measure of the best that could be achieved using the KR method, that is, an upper bound on its performance. The estimated results are marked with an asterisk in the table. Throughout Table IV, the KR adjustment is seen to give a test statistic with inflated size (for both $T = 5$ and 10 time points), although this size is seen to approach the nominal level of 5% as the number of subjects increases.

In the extended simulations of study designs (A′) and (B′) tests involving the adjusted sandwich estimator are seen to control size, but these tests appear to have little power to detect differences from the null hypothesis in comparison to the other methods. This is particularly true for data arising from underlying covariance structures which are far from the independence (identity) 'working' covariance structure. Also, for increased time points ($T = 10$), there were problems with this adjustment in the low subject setting ($n = 20$),

Table III. Summary of results from 1000 simulations of extended study design (A′).

Underlying 'true' covariance structure	Number of time points (T) and subjects (n)		Proportion of significant test results (out of 1000) (Null model - no treatment/time interaction)				
			Method of inf.	Unstr. (KR)	Sand. adj. (PW)	Box F	Mod. Box
Stationary structures							
Compound symmetry	T=5	n=10	Size	0.045	0.057	0.023	0.042
			Power	0.747	0.660	0.964	0.984
		n=20	Size	0.057	0.037	0.024	0.036
			Power	0.749	0.663	0.808	0.843
	T=10	n=20	Size	0.049	0.035	0.013	0.020
			Power	0.756	0.510	0.950	0.964
		n=40	Size	0.049	0.034	0.031	0.038
			Power	0.756	0.645	0.827	0.840
AR1 (ρ=0.8)	T=5	n=10	Size	0.056	0.083	0.035	0.068
			Power	0.775	0.793	0.992	0.999
		n=20	Size	0.042	0.037	0.032	0.052
			Power	0.756	0.884	0.937	0.958
	T=10	n=20	Size	0.048	0.065	0.038	0.060
			Power	0.753	0.698	0.995	0.996
		n=40	Size	0.052	0.048	0.037	0.053
			Power	0.779	0.728	0.968	0.972
Non-stationary structures							
Heterogeneous comp. sym.	T=5	n=10	Size	0.053	0.044	0.026	0.052
			Power	0.773	0.534	0.961	0.981
		n=20	Size	0.042	0.017	0.033	0.052
			Power	0.740	0.586	0.788	0.823
	T=10	n=20	Size	0.052	0.016	0.026	0.040
			Power	0.758	0.207	0.980	0.989
		n=40	Size	0.063	0.023	0.023	0.028
			Power	0.765	0.550	0.901	0.910
Heterogeneous AR1 (ρ=0.8)	T=5	n=10	'Size'	0.049	0.076	0.034	0.069
			Power	0.744	0.765	0.981	0.994
		n=20	Size	0.043	0.046	0.037	0.053
			Power	0.770	0.721	0.922	0.947
	T=10	n=20	Size	0.051	0.042	0.035	0.054
			Power	0.767	0.604	0.990	0.996
		n=40	Size	0.053	0.044	0.031	0.047
			Power	0.755	0.687	0.912	0.932
Antedependence	T=5	n=10	Size	0.060	0.053	0.038	0.059
			Power	0.767	0.611	0.861	0.924
		n=20	Size	0.058	0.032	0.041	0.053
			Power	0.741	0.600	0.624	0.688
	T=10	n=20	Size	0.043	0.003	0.041	0.052
			Power	0.749	0.070	0.834	0.861
		n=40	Size	0.056	0.010	0.038	0.043
			Power	0.764	0.403	0.704	0.725

Design(A′): a simple repeated measures experiment with n subjects in two treatment groups of equal size, and measurements recorded on T occasions. Table gives the proportion of type 1 errors (Size) and Power assessed using a fixed level of $\alpha = 0.05$. Power is calculated by adding a linear effect in time to one treatment group to achieve approximately 75% using the Kenward-Roger adjustment with an unstructured covariance matrix. KR - Kenward-Roger adjusted test; PW - Pan and Wall bias-adjusted sandwich estimator. The Box and modified Box correction statistics (Box F and Mod Box) adopt the unstructured REML estimator of the covariance matrix.

Table IV. **Summary of results from 1000 simulations of extended study design (B′).**

Underlying 'true' covariance structure	Number of time points (T) and subjects (n)		Method of inf.	Proportion of significant test results (out of 1000) (Null model - no treatment/time interaction)			
				Unstr. (KR)	Sand. adj. (PW)	Box F	Mod. Box
Stationary structures							
Compound symmetry	$T=5$	$n=10$	Size	0.074	0.051	0.017	0.040
		($miss=2$)	Power	0.632	0.523	0.882	0.937
		$n=20$	Size	0.059	0.036	0.041	0.050
		($miss=4$)	Power	0.644	0.505	0.733	0.772
	$T=10$	$n=20$	Size	0.053*	0.050	0.012	0.023
		($miss=4$)	Power	0.624*	0.361	0.857	0.890
		$n=40$	Size	0.057*	0.038	0.046	0.052
		($miss=8$)	Power	0.624*	0.449	0.610	0.643
AR1 ($\rho=0.8$)	$T=5$	$n=10$	Size	0.078	0.064	0.024	0.061
		($miss=2$)	Power	0.637	0.677	0.966	0.989
		$n=20$	Size	0.067	0.050	0.052	0.067
		($miss=4$)	Power	0.662	0.586	0.869	0.900
	$T=10$	$n=20$	Size	0.068*	0.096	0.036	0.068
		($miss=4$)	Power	0.645*	0.618	0.993	0.995
		$n=40$	Size	0.062*	0.040	0.054	0.070
		($miss=8$)	Power	0.674*	0.614	0.925	0.944
Non-stationary structures							
Heterogeneous comp. sym.	$T=5$	$n=10$	Size	0.084	0.056	0.019	0.042
		($miss=2$)	Power	0.609	0.412	0.843	0.927
		$n=20$	Size	0.054	0.032	0.043	0.059
		($miss=4$)	Power	0.634	0.439	0.681	0.739
	$T=10$	$n=20$	Size	0.062*	0.025	0.022	0.037
		($miss=4$)	Power	0.627*	0.162	0.910	0.939
		$n=40$	Size	0.063*	0.020	0.030	0.038
		($miss=8$)	Power	0.667*	0.318	0.796	0.810
Heterogeneous AR1 ($\rho=0.8$)	$T=5$	$n=10$	Size	0.083	0.066	0.031	0.072
		($miss=2$)	Power	0.631	0.578	0.914	0.975
		$n=20$	Size	0.048	0.025	0.052	0.070
		($miss=4$)	Power	0.683	0.562	0.836	0.884
	$T=10$	$n=20$	Size	0.048*	0.049	0.048	0.075
		($miss=4$)	Power	0.602*	0.413	0.948	0.974
		$n=40$	Size	0.058*	0.038	0.042	0.054
		($miss=8$)	Power	0.658*	0.505	0.851	0.894
Antedependence	$T=5$	$n=10$	Size	0.065	0.042	0.038	0.072
		($miss=2$)	Power	0.664	0.494	0.748	0.862
		$n=20$	Size	0.054	0.028	0.040	0.057
		($miss=4$)	Power	0.658	0.472	0.558	0.634
	$T=10$	$n=20$	Size	0.071*	0.000	0.034	0.049
		($miss=4$)	Power	0.604*	0.040	0.767	0.813
		$n=40$	Size	0.048*	0.005	0.036	0.043
		($miss=8$)	Power	0.654*	0.174	0.603	0.628

Design(B′): as Design(A′) but with ($miss$) subjects allowed to drop out at random following the first observation as indicated. Table gives the proportion of type 1 errors (Size) and Power assessed using a fixed level of $\alpha = 0.05$. Power is calculated by adding a linear effect in time to one treatment group to achieve approximately 75% using the Kenward-Roger adjustment with an unstructured covariance matrix where the data are complete. KR - Kenward-Roger adjusted test; PW - Pan and Wall bias-adjusted sandwich estimator. The Box and modified Box correction statistics (Box F and Mod Box) adopt the unstructured REML estimator of the covariance matrix. The asterisked values (*) are estimated using 'known' underlying covariance structures and the average number of observations.

Table V. Summary of results from 1000 simulations of extended study designs (C′) and (D).

Underlying 'true' covariance structure	Number of treatment/ time points (T) and subjects (n)		Method of Inf.	Proportion of significant test results (out of 1000) (Null model - no treatment effect)			
				Unstr. (KR)	Sand. adj. (PW)	Box F	Mod. Box
Stationary Structures							
Compound symmetry	$T=5$	$n=10$	Size	0.690	0.034	0.032	0.046
			Power	——	0.564	0.958	0.977
		$n=20$	Size	0.126	0.045	0.049	0.062
			Power	——	0.900	0.981	0.983
	$T=9$	$n=18$	Size	0.472*	0.037	0.033	0.040
			Power	——	0.531	0.991	0.994
		$n=36$	Size	0.063*	0.029	0.041	0.047
			Power	0.977*	0.941	0.997	0.997
AR1 (ρ=0.8)	$T=5$	$n=10$	Size	0.672	0.034	0.020	0.043
			Power	——	0.313	0.738	0.827
		$n=20$	Size	0.123	0.024	0.036	0.053
			Power	——	0.533	0.837	0.867
	$T=9$	$n=18$	Size	0.497*	0.020	0.020	0.033
			Power	——	0.139	0.656	0.706
		$n=36$	Size	0.086*	0.023	0.037	0.052
			Power	0.985*	0.406	0.778	0.796
Non-stationary structures							
Heterogeneous comp. sym.	$T=5$	$n=10$	Size	0.682	0.032	0.016	0.047
			Power	——	0.372	0.840	0.899
		$n=20$	Size	0.124	0.033	0.038	0.053
			Power	——	0.691	0.904	0.926
	$T=9$	$n=18$	Size	0.499*	0.016	0.028	0.037
			Power	——	0.300	0.864	0.892
		$n=36$	Size	0.065*	0.032	0.044	0.048
			Power	0.974*	0.671	0.922	0.929
Heterogeneous AR1 (ρ=0.8)	$T=5$	$n=10$	Size	0.597	0.037	0.012	0.028
			Power	——	0.226	0.442	0.536
		$n=20$	Size	0.111	0.024	0.029	0.043
			Power	——	0.370	0.665	0.717
	$T=9$	$n=18$	Size	0.473*	0.023	0.016	0.030
			Power	——	0.100	0.471	0.535
		$n=36$	Size	0.069*	0.026	0.036	0.046
			Power	0.987*	0.237	0.591	0.619
Antedependence	$T=5$	$n=10$	Size	0.697	0.028	0.024	0.037
			Power	——	0.206	0.618	0.698
		$n=20$	Size	0.122	0.028	0.048	0.054
			Power	——	0.398	0.688	0.712
	$T=9$	$n=18$	Size	0.483*	0.018	0.029	0.023
			Power	——	0.126	0.515	0.403
		$n=36$	Size	0.066*	0.016	0.044	0.051
			Power	0.951*	0.242	0.579	0.592

Designs(C′) and (D): cross-over studies with $T = 5$ or 9 treatments and periods, and n subjects assigned to treatments using Williams' squares. Table gives the proportion of type 1 errors (Size) and Power assessed using a fixed level of α =0.05. Power is calculated by adding a linear effect increasing by treatment to all subjects. For comparison purposes the treatment effect is calculated by reference to a Kenward-Roger adjustment using the known covariance matrix to achieve approximately 100%. KR - Kenward-Roger adjusted test; PW - Pan and Wall bias-adjusted sandwich estimator. The Box and modified Box correction statistics (Box F and Mod Box) adopt the unstructured REML estimator of the covariance matrix. The asterisked values (*) are estimated using 'known' underlying covariance structures. Power is not assessed where the test size is not fixed at the nominal level (—).

resulting in invalid (negative) estimates for the denominator degrees of freedom. In these instances, setting the numerator degrees of freedom to c, the dimensionality of the test, results in a single denominator degree of freedom test. These issues do not, however, reccur as the number of subjects increases, and the results for 10 time points and 20 subjects do not appear to be out of line.

Consider now the cross-over studies of designs (C′) and (D). For extended study design (C′), based on the five treatment-five period design, we see that as the number of subjects increases from 10 to 20, the test sizes using the KR adjustment are closer to the nominal level of 5%, but are too inflated for power to be considered. Power is compared for the adjusted sandwich estimator and the Box corrections for treatment differences which lead to comparable powers using the Wald statistic (with a KR adjustment and a 'known' covariance structure) of 100%. Again, the modified Box correction is seen to give a test with nominal properties and good power in comparison to the 'true' test.

Results from the simulations under study design (D) are obtained similarly, and the modified Box correction is again seen to give the better performance. As previously seen, the adjusted sandwich estimator results in invalid parameter estimates where the number of time points is increased and the number of subjects is small ($n = 18$), so that the denominator degrees of freedom of such tests are fixed at 1, close to the boundary.

In this setting, with an estimated (9×9) covariance structure, the KR adjustments are again computationally expensive in terms of 1000 individual simulations, so are estimated, for comparison purposes, using the 'known' structure to give an upper bound on the performance. (Again these estimated values are marked with an asterisk in the table). These show that the test size reduces towards nominal levels as the number of subjects increases, so that it may be appropriate to consider power. However, it is clear that these values cannot be achieved in practice.

6. Examples

Two examples are presented which illustrate the use of the modified Box correction in practical analyses.

6.1. Cardiac enzyme in preserved dog hearts

Data were collected to investigate the preserving quality of a liquid on the enzyme content (%ATP) of dog hearts. The data presented concern 12 dog hearts, and differ according to the presence or absence of a component, A, in the preserving liquid. (The data are a subset of a larger factorial design to compare two components, A and B, and correspond to the presence or absence of A, B absent). The mean profiles of the two treatment groups are
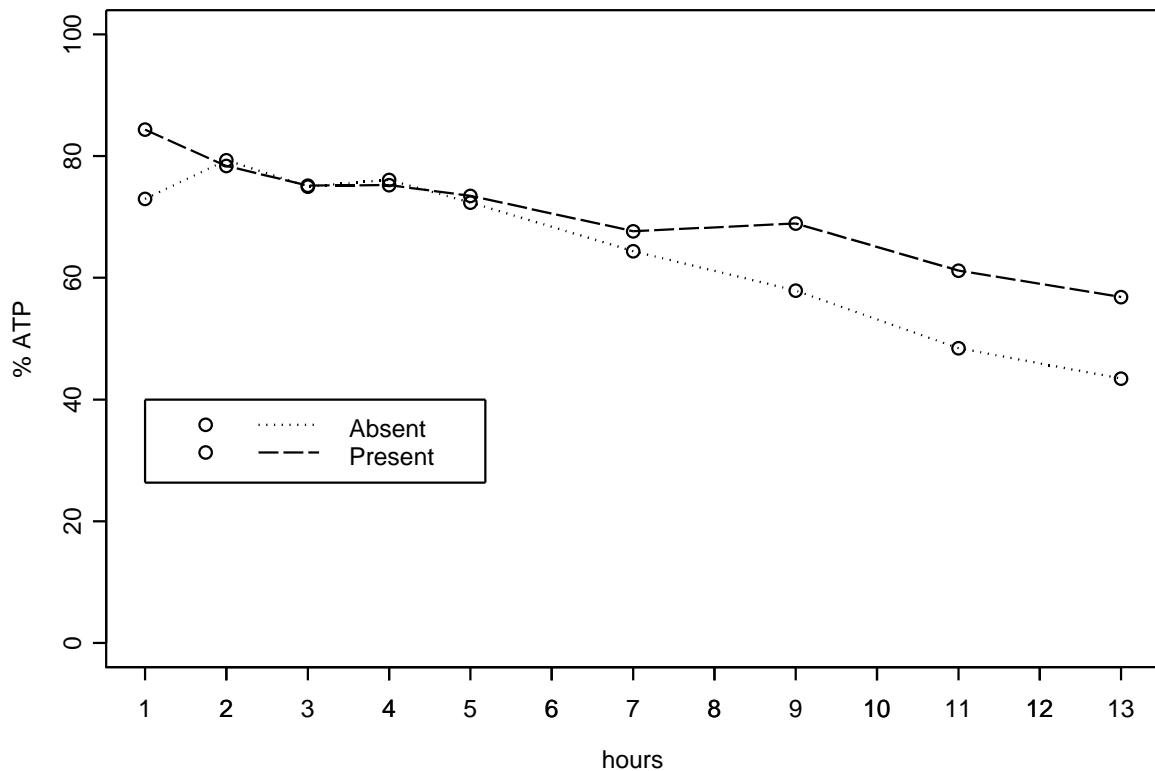
Figure 1. **Cardiac enzyme data: mean profiles.**

shown as Figure 1 and the pooled within-groups sample covariance-correlation matrix as Table VI.

Treating these data as a simple repeated measures design, a test for a treatment by time interaction using the unstructured covariance matrix with a Kenward-Roger (KR) small sample adjustment is equivalent to an exact Hotelling T^2 test. The upper panel of Table VII shows the results obtained from Wald tests using various covariance models and the modified Box correction. It can be seen that the test results differ according to the choice of structure. As the data are complete and balanced, the same mean estimates (by ordinary least squares) are obtained under each method, but they differ in their estimates of precision.

The exact Hotelling T^2 test obtained using the unstructured covariance matrix indicates that there is insufficient evidence (at the 5% level) to reject the null hypothesis of no treatment by time interaction. This is confirmed by the modified Box corrected statistic, although in this instance the evidence is less marginal than that offered by the Hotelling T^2 test.

To show how such results can differ where there is imbalance, *i.e.* when the estimates of the

Table VI. Sample covariance-correlation matrix for the cardiac enzyme data.

37.08	11.29	4.04	32.53	24.78	37.22	51.32	19.08	15.89
0.34	**29.27**	-3.52	12.80	7.64	10.02	18.66	8.14	-7.84
0.12	-0.11	**33.08**	-7.70	15.43	6.91	15.80	-11.43	30.00
0.47	0.21	-0.12	**128.08**	-27.86	6.51	58.84	19.38	-43.17
0.58	0.20	0.38	-0.35	**48.85**	46.33	33.20	24.45	53.01
0.57	0.17	0.11	0.05	0.62	**114.22**	86.48	44.59	61.27
0.78	0.32	0.25	0.48	0.44	0.75	**117.38**	51.39	48.76
0.30	0.14	-0.19	0.16	0.33	0.40	0.45	**111.24**	42.10
0.27	-0.15	0.54	-0.39	0.78	0.59	0.46	0.41	**94.24**

Variances on the diagonal with covariances above and correlations below.

Table VII. Cardiac enzyme data: comparison of results.

Covariance structure	Num. df	Den. df	F	p
Complete data				
Identity - independence	8	90	1.49	0.1713
Unstructured	8	3	8.73	0.0509
Compound symmetry	8	80	2.07	0.0485
AR1	8	73.8	1.24	0.2904
Mod. Box ($\lambda = 0.59$)	8	11.2	2.52	0.0774
With artificial dropout				
Identity - independence	8	87	1.87	0.0753
Unstructured	8	1.6	88.63	0.0252
Compound symmetry	8	77.2	2.32	0.0274
AR1	8	12.2	1.51	0.1686
Mod. Box ($\lambda = 0.66$)	8	10.48	2.84	0.0591

Table gives results of Wald tests using a Kenward-Roger adjustment, under the given covariance model assumptions, for a treatment/time interaction. The comparative results using the modified Box correction adopt the unstructured REML estimator of the covariance matrix.

mean parameters as well as their standard errors are dependent on the choice of covariance structure, we introduce an artificial dropout to this reduced data set. To achieve this, consider that the three final measurements are missing from one of the dog hearts which receives the preserving liquid from which component A is absent. Repeating the tests for a treatment/time interaction with this artificial dropout gives the results shown in the lower panel of Table VII.

Now the tests using the KR adjustment (with the identity, unstructured, and compound symmetry structures) are no longer exact, and so these results are less plausible given the performance of such tests in our simulations. There is a higher significance of an interaction using the unstructured form once the three observations from the 'absent' group are removed, but the test using the modified Box correction remains non-significant.

6.2. Electrocardiogram abnormalities in the guinea pig papillary muscle

As the second example, we present analysis of data published by Brammer [23] which illustrated the need for methods specific to very small samples of repeated measurements.

These data comprise measurements taken from papillary muscles dissected from the right ventricles of each of just three guinea pigs' hearts in two experiments. The purpose of the experiments was to determine whether the compounds are likely to cause electrocardiogram abnormalities. Brammer recognises that analysis from such small samples is unlikely to be definitive, but notes that such small samples are common in isolated tissue or organ experiments.

Since the isolated tissue assays from the guinea pigs deteriorate in time, there is a limited period in which to test different concentrations of the compounds on each muscle, so a control measure is followed by six increasing concentrations of the compound. In such an ascending dose design, the carryover effect is considered to be minimal in comparison to the current dose. Concentration and time are confounded, but a separate 'control' experiment with no compound present showed that there were no important changes over time. Five variables were measured, but we focus here on AP (amplitude of action potential). Mean profiles under each compound are shown in Figure 3.

These experiments can be considered as block designs, with concentration of compound as the treatment and tissue as the blocking factor, but the compound symmetry structure imposed by such a design may not be appropriate. Instead, Brammer treats the experiments as simple repeated measures designs with concentration as the time variable and tissue as the subject, and compares the resulting analyses from adopting various covariance models to account for the correlation between measurements on the same tissue (subject).

To analyse these experiments in such a way requires the fitting of a (7×7) covariance structure using just three subjects, so unsurprisingly the estimation method for the unstructured
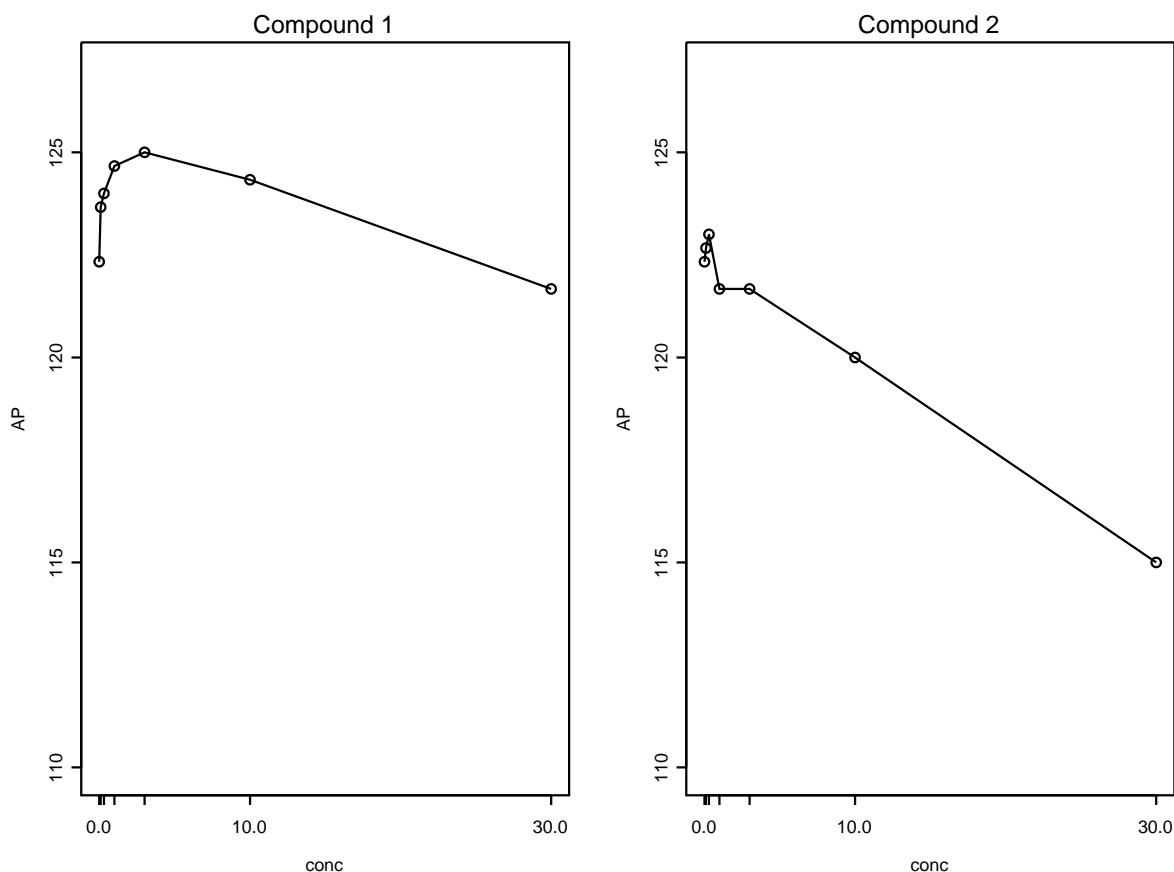
Figure 2. GPPM data: mean profiles, compounds 1 and 2.

covariance models did not converge. Brammer compares those correlation structures that can be fitted by informal comparison of the (reduced) log-likelihoods and prefers an AR1 or heterogeneous AR1 model in each case over the more usual compound symmetry approach adopted for such experiments. However, the results of the simulations in [1] show that, for small samples, such methods are unreliable for choosing an appropriate structure.

It is of interest to compare Brammer's approach with that offered by the modified Box correction in this extreme small sample setting. Whilst the estimation method for the unstructured covariance model did not converge for either of the two experiments, it is possible to construct the (singular) sample covariance matrix directly in each case. The resulting estimates are shown in Table VIII below. While these matrices do not allow the construction of the usual Wald tests, which require invertible matrices, they can be used directly in the modified Box correction to allow the tests to reflect the observed dependencies in the data.

Although it is possible to replicate Brammer's tests by suitable parametrization of the design matrix, with columns representing contrasts between each concentration with the control

17

Table VIII. Sample covariance-correlation matrices for the GPPM data: compounds 1 and 2.

Compound 1

2.33	3.67	3.00	4.17	3.50	3.83	7.17
0.95	**6.33**	5.00	7.83	6.50	6.17	12.83
0.98	0.99	**4.00**	6.00	5.00	5.00	10.00
0.85	0.97	0.93	**10.33**	8.50	7.17	16.33
0.87	0.98	0.94	1.00	**7.00**	6.00	13.50
1.00	0.97	0.99	0.89	0.90	**6.33**	12.17
0.91	0.99	0.97	0.99	0.99	0.94	**26.33**

Compound 2

21.33	22.67	24.00	22.67	22.67	20.00	20.00
0.99	**24.33**	25.50	24.83	24.83	22.50	23.00
1.00	0.99	**27.00**	25.50	25.50	22.50	22.50
0.96	0.98	0.96	**26.33**	26.33	25.00	26.50
0.96	0.98	0.96	1.00	**26.33**	25.00	26.50
0.87	0.91	0.87	0.97	0.97	**25.00**	27.50
0.78	0.84	0.78	0.93	0.93	0.99	**31.00**

Variances on the diagonal with covariances above and correlations below.

measurement, a more appropriate method is suggested as follows.

1. Use the modified Box correction to test for an overall treatment (concentration) effect.

2. If significant, use Scheffé's method, in conjunction with the adjusted F-statistic, to test individual contrasts.

This approach ensures that the type 1 error rate for individual tests is controlled for multiplicity of testing, as well as to departures from independence in the small sample setting for which the modified Box correction has been shown to be successful for the analysis of repeated measurements.

Scheffé's method [11] allows for the comparison of any or all possible contrasts between treatment means, ensuring that the type 1 error rate is at most α for any of the possible comparisons. It takes advantage of the union-intersection test properties of the ANOVA F-statistic, by simultaneously considering all possible contrasts in the treatment means:

$$\Gamma_{\mathbf{a}} = a_1\mu_1 + a_2\mu_2 + \ldots + a_r\mu_r \tag{21}$$

for any $\mathbf{a}=(a_1,\ldots,a_r)$, with $\sum a_i = 0$. The corresponding contrasts in the treatment averages $\bar{y}_{i.}$, are hence

$$C_{\mathbf{a}} = a_1\bar{y}_{1.} + a_2\bar{y}_{2.} + \ldots + a_r\bar{y}_{r.} \tag{22}$$

and the standard error of this contrast is

$$S_{C_{\mathbf{a}}} = \sqrt{\hat{\sigma}^2 \sum_{i=1}^{r}(a_i^2/n_i)} \tag{23}$$

where n_i is the number of observations of the ith treatment, and $\hat{\sigma}^2$ is the mean squared error (MSE) of the data.

To use Scheffé's method with the modified Box corrected ANOVA statistic from (12) with (13)-(15), we have $\hat{\sigma}^2 = \text{MSE} = \mathbf{y}^T \mathbf{A} \mathbf{y}/(n-r)$, and the critical value to which $C_{\mathbf{a}}$ should be compared is

$$S_{\alpha,\mathbf{a}} = S_{C_{\mathbf{a}}} \sqrt{\frac{c}{\lambda} F_{\alpha;c,v_2}} \tag{24}$$

so that a 100(1-α)% confidence interval for $C_{\mathbf{a}}$ is given by

$$C_{\mathbf{a}} \pm S_{\alpha,\mathbf{a}} = C_{\mathbf{a}} \pm S_{C_{\mathbf{a}}} \sqrt{\frac{c}{\lambda} F_{\alpha;c,v_2}} \tag{25}$$

An advantage of Scheffé's method is that it will always agree with the ANOVA F-test in the sense that if the F-test detects differences, then at least one Scheffé test will detect a difference. Conversely, if the F-test does not detect any differences, then none of Scheffé's tests will. This is illustrated below using Brammer's data.

We begin by considering the overall tests for a concentration effect, given in Table IX. The results show that there is only marginal evidence of a significant effect of concentration with compound 1, but that the evidence of a significant effect in compound 2 is much stronger.

Table IX. **GPPM data: results using the modified Box correction.**

Effect		Num df	Den df	F	p
Concentration (compound 1)	(λ=0.1131)	6	5.1756	4.4977	0.0570
Concentration (compound 2)	(λ=0.0429)	6	5.0861	20.8472	0.0020

Table gives results of tests of overall treatment effect using the modified Box correction adopting the *singular* sample covariance matrix.

Turning to the Scheffé tests of individual contrasts (differences from control), we have, for compound 1, $\hat{\sigma}^2$=8.9524, so that the standard error for the first, and actually all, the contrasts is given by

$$S_{C_1} = \sqrt{\hat{\sigma}^2 \sum_{i=1}^{7} (a_i^2/n_i)} = \sqrt{8.9524(1+1)/3} = 2.4430$$

and, since the 0.95-quantile of F(6,5.1756) is 4.8063, we find the 95% confidence interval for the first contrast is

$$1.333 \pm 2.4430\sqrt{6 \times 0.1131 \times 4.8063} = 1.333 \pm 4.4120 = (-3.08, 5.75)$$

Table X. GPPM data: individual contrasts using Scheffé's method with the modified Box correction.

	Compound 1		Compound 2	
	Mean diff.		Mean diff.	
Conc	from control	95% CI	from control	95% CI
0				
0.1	1.3333	(-3.08,5.74)	0.3333	(-4.33,4.99)
0.3	1.6667	(-2.74,6.08)	0.6667	(-3.99,5.33)
1	2.3333	(-2.07,6.75)	-0.6667	(-5.33,3.99)
3	2.6667	(-1.74,7.08)	-0.6667	(-5.33,3.99)
10	2.0000	(-2.71,6.41)	2.3333	(-6.99,2.33)
30	-0.6667	(-5.07,3.74)	-7.3333	(-11.99,-2.67)

Table gives Scheffé confidence intervals for individual concentration differences from control using the modified Box statistics of Table IX.

Confidence intervals for the remaining contrasts are calculated similarly, (for compound 2, $\hat{\sigma}^2 = 25.9048$). The results are shown in Table X.

As is expected, since the overall concentration effect was non-significant at the 5% level, none of the 95% confidence intervals for mean difference from control for compound 1 exclude zero. Of the contrasts with compound 2, only the final concentration is significantly different from control, 95% CI (-11.99, 2.67). (In fact this contrast is also significant at the 1% level, 99% CI (-14.14, -0.52)).

7. Discussion

The extensive simulation studies of Section 5 show that the modified Box correction results in a test with correct size which is more powerful than the other methods considered across a range of small sample settings for the analysis of repeated measurements. This is also seen to be the case in further simulations undertaken by the authors, for data arising from underlying covariance structures with low to medium correlation, such as the independence and AR1 ($\rho = 0.2$) structures used in the simulations of the companion paper. For such data, as might be expected, the Box corrections are most powerful, as we are closest in this setting to the underlying assumptions of the ANOVA statistic from which they originate.

Tests based on the adjusted sandwich estimator are also seen to have correct size across the range of settings considered, but this approach lacks power. Also, estimation of the adjusted test parameters are seen to be non-robust, resulting in invalid estimates of the denominator degrees of freedom where the number of time points is large in comparison to the number of subjects. It is worth noting that the adjusted sandwich estimator method of Pan and Wall does not assume a common covariance structure for all subjects, and that this may contribute to its loss of power in comparison to the Box corrections. However, we believe that the true advantage of the Box corrections relates not to how the covariance matrix is assessed across

20

subjects, but rather how it is used in inference. That is, where exactly such assumptions about the covariance structure enter the process. Further simulations (not presented) have shown no significant loss of power using the Box corrections where a common covariance structure is not assumed, such as when allowing the structure to change between treatment groups. However, convergence problems can arise in estimating the unstructured form using REML in this instance in the most extreme small sample settings.

The simulations confirm that Wald tests using an unstructured covariance matrix with the Kenward-Roger adjustment give inflated type 1 error rates where the data do not allow for exact tests, although the size of such tests does approach nominal levels as the sample size increases as we might expect. Where nominal properties are achieved, so that it is appropriate to consider power, we see that where the sample size (number of subjects) is small, tests using the Box corrections give greater power than the corresponding Wald tests. As the sample size increases the improvement in power from using the Box corrections over the KR adjustment diminishes as the underlying covariance structure moves further from independence, but the correction remains an effective method for inference.

The modified Box correction developed in Section 3 is preferred to Box's original statistic which is conservative and hence less powerful. This method can, by suitable parameterization of the design matrix, be used to test any hypothesis involving fixed effects, based on their ordinary least squares estimates under the assumption of independence, and using any consistent estimator of the covariance matrix, such as that obtained from REML. It can be easily implemented using statistical software with minimal programming. Also, seen in the examples in Section 6 it is easily combined with Scheffé's method for simultaneous contrasts to examine questions of interest arising from significant tests, and providing appropriate control for multiple testing.

Appendix A

The following (5×5) and (10×10) symmetric matrices are used as the underlying covariance structures for the data generated in the simulations of extended designs (A$'$), (B$'$)and (C$'$) used in the Section 5. The (9×9) matrices required for study design (D), are simply extracted from the first nine rows and columns of the (10×10) structures noted above.

A.1. Compound Symmetry

$$
\begin{bmatrix}
\sigma_1^2 + \sigma^2 & & & & & \\
\sigma_1^2 & \sigma_1^2 + \sigma^2 & & & & \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 & & & \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 & & \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 + \sigma^2
\end{bmatrix}
$$

with $\sigma_1^2 = 1$ and $\sigma^2 = 1$. We obtain the (5×5) and (10×10) matrices

$$
\begin{bmatrix}
2 & & & & \\
1 & 2 & & & \\
1 & 1 & 2 & & \\
1 & 1 & 1 & 2 & \\
1 & 1 & 1 & 1 & 2
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
2 & & & & & & & & & \\
1 & 2 & & & & & & & & \\
1 & 1 & 2 & & & & & & & \\
1 & 1 & 1 & 2 & & & & & & \\
1 & 1 & 1 & 1 & 2 & & & & & \\
1 & 1 & 1 & 1 & 1 & 2 & & & & \\
1 & 1 & 1 & 1 & 1 & 1 & 2 & & & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2
\end{bmatrix}
$$

A.2. AR1

$$
\sigma^2
\begin{bmatrix}
1 & & & & & \\
\rho & 1 & & & & \\
\rho^2 & \rho & 1 & & & \\
\rho^3 & \rho^2 & \rho & 1 & & \\
\rho^4 & \rho^3 & \rho^2 & \rho & 1 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\rho^9 & \rho^8 & \rho^7 & \rho^6 & \rho^5 & \dots & \rho
\end{bmatrix}
$$

with $\sigma^2 = 1$ and $\rho = 0.8$. We obtain the (5×5) and (10×10) matrices

$$
\begin{bmatrix}
1 & & & & \\
0.8 & 1 & & & \\
0.64 & 0.8 & 1 & & \\
0.512 & 0.64 & 0.8 & 1 & \\
0.4096 & 0.512 & 0.64 & 0.8 & 1
\end{bmatrix}
$$

and

$$\begin{bmatrix}
1 \\
0.8 & 1 \\
0.64 & 0.8 & 1 \\
0.512 & 0.64 & 0.8 & 1 \\
0.4096 & 0.512 & 0.64 & 0.8 & 1 \\
0.3277 & 0.4096 & 0.512 & 0.64 & 0.8 & 1 \\
0.2622 & 0.3277 & 0.4096 & 0.512 & 0.64 & 0.8 & 1 \\
0.2098 & 0.2622 & 0.3277 & 0.4096 & 0.512 & 0.64 & 0.8 & 1 \\
0.1678 & 0.2098 & 0.2622 & 0.3277 & 0.4096 & 0.512 & 0.64 & 0.8 & 1 \\
0.1343 & 0.1678 & 0.2098 & 0.2622 & 0.3277 & 0.4096 & 0.512 & 0.64 & 0.8 & 1
\end{bmatrix}$$

A.3. Heterogeneous compound symmetry

$$\begin{bmatrix}
\sigma_1^2 \\
\sigma_2\sigma_1\rho & \sigma_2^2 \\
\sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 \\
\sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \\
\sigma_5\sigma_1\rho & \sigma_5\sigma_2\rho & \sigma_5\sigma_3\rho & \sigma_5\sigma_4\rho & \sigma_5^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\sigma_{10}\sigma_1\rho & \sigma_{10}\sigma_2\rho & \sigma_{10}\sigma_3\rho & \sigma_{10}\sigma_4\rho & \sigma_{10}\sigma_5\rho & \cdots & \sigma_{10}^2
\end{bmatrix}$$

For the (5×5) structure, we use $\rho = 0.5$, with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, and $\sigma_5^2 = 5$, to obtain

$$\begin{bmatrix}
1 \\
0.7071 & 2 \\
0.8660 & 1.2247 & 3 \\
1.0000 & 1.4142 & 1.7321 & 4 \\
1.1180 & 1.5811 & 1.9365 & 2.2361 & 5
\end{bmatrix}$$

For the (10×10) structure, we use $\rho = 0.5$, with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, $\sigma_5^2 = 5$,

$\sigma_6^2 = 6$, $\sigma_7^2 = 7$, $\sigma_8^2 = 8$, $\sigma_9^2 = 9$ and $\sigma_{10}^2 = 10$, to obtain

$$
\begin{bmatrix}
1 \\
0.7071 & 2 \\
0.8660 & 1.2247 & 3 \\
1.0000 & 1.4142 & 1.7321 & 4 \\
1.1180 & 1.5811 & 1.9365 & 2.2361 & 5 \\
1.2247 & 1.7321 & 2.1213 & 2.4495 & 2.7386 & 6 \\
1.3229 & 1.8708 & 2.2913 & 2.6458 & 2.9580 & 3.2404 & 7 \\
1.4142 & 2.0000 & 2.4495 & 2.8284 & 3.1623 & 3.4641 & 3.7417 & 8 \\
1.5000 & 2.1213 & 2.5981 & 3.0000 & 3.3541 & 3.6742 & 3.9686 & 4.2426 & 9 \\
1.5811 & 0.2361 & 2.7386 & 3.1623 & 3.5355 & 3.8730 & 4.1833 & 4.4721 & 4.7434 & 10
\end{bmatrix}
$$

A.4. Heterogeneous AR1

$$
\begin{bmatrix}
\sigma_1^2 \\
\sigma_2\sigma_1\rho & \sigma_2^2 \\
\sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 \\
\sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \\
\sigma_5\sigma_1\rho^4 & \sigma_5\sigma_2\rho^3 & \sigma_5\sigma_3\rho^2 & \sigma_5\sigma_4\rho & \sigma_5^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\sigma_{10}\sigma_1\rho^9 & \sigma_{10}\sigma_2\rho^8 & \sigma_{10}\sigma_3\rho^7 & \sigma_{10}\sigma_4\rho^6 & \sigma_{10}\sigma_5\rho^5 & \cdots & \sigma_{10}^2
\end{bmatrix}
$$

For the (5×5) structure, we use $\rho = 0.8$, with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, and $\sigma_5^2 = 5$, to obtain

$$
\begin{bmatrix}
1 \\
0.1314 & 2 \\
1.1085 & 1.9596 & 3 \\
1.0240 & 1.8102 & 2.7713 & 4 \\
0.9159 & 1.6191 & 2.4787 & 3.5777 & 5
\end{bmatrix}
$$

For the (10×10) structure, we use $\rho = 0.8$, with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, $\sigma_5^2 = 5$, $\sigma_6^2 = 6$, $\sigma_7^2 = 7$, $\sigma_8^2 = 8$, $\sigma_9^2 = 9$ and $\sigma_{10}^2 = 10$, to obtain

$$
\begin{bmatrix}
1 & & & & & & & & & \\
0.1314 & 2 & & & & & & & & \\
1.1085 & 1.9596 & 3 & & & & & & & \\
1.0240 & 1.8102 & 2.7713 & 4 & & & & & & \\
0.9159 & 1.6191 & 2.4787 & 3.5777 & 5 & & & & & \\
0.8026 & 1.4189 & 2.1722 & 3.1353 & 4.3818 & 6 & & & & \\
0.6936 & 1.2261 & 1.8770 & 2.7092 & 3.7863 & 5.1846 & 7 & & & \\
0.5932 & 1.0486 & 1.6053 & 2.3170 & 3.2382 & 4.4341 & 5.9867 & 8 & & \\
0.5033 & 0.8897 & 1.3621 & 1.9661 & 2.7477 & 3.7624 & 5.0798 & 6.7882 & 9 & \\
0.4244 & 0.7503 & 1.1487 & 1.6579 & 2.3170 & 3.1727 & 4.2837 & 5.7243 & 7.5895 & 10
\end{bmatrix}
$$

A.5. Antedependence

$$
\begin{bmatrix}
\sigma_1^2 & & & & & \\
\sigma_1\sigma_2\rho_1 & \sigma_2^2 & & & & \\
\sigma_1\sigma_3\rho_1\rho_2 & \sigma_2\sigma_3\rho_2 & \sigma_3^2 & & & \\
\sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_2\sigma_4\rho_2\rho_3 & \sigma_3\sigma_4\rho_3 & \sigma_4^2 & & \\
\sigma_1\sigma_5\rho_1\rho_2\rho_3\rho_4 & \sigma_2\sigma_5\rho_2\rho_3\rho_4 & \sigma_3\sigma_5\rho_3\rho_4 & \sigma_4\sigma_5\rho_4 & \sigma_5^2 & \\
\vdots & & & & & \ddots \\
\sigma_1\sigma_{10}\rho 1\rho_2\rho_3\rho_4\rho_5\rho_6\rho_7\rho_8\rho_9 & & & & & \sigma_{10}^2
\end{bmatrix}
$$

With $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, $\sigma_5^2 = 5$, $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_3 = 0.4$, and $\rho_4 = 0.2$, we have the (5×5) matrix

$$
\begin{bmatrix}
1 & & & & \\
1.1314 & 2 & & & \\
0.8314 & 1.4697 & 3 & & \\
0.3840 & 0.6788 & 1.3856 & 4 & \\
0.0859 & 0.1518 & 0.3098 & 0.8944 & 5
\end{bmatrix}
$$

For the (10×10) structure, we use $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 3$, $\sigma_4^2 = 4$, $\sigma_5^2 = 5$, $\sigma_6^2 = 6$, $\sigma_7^2 = 7$, $\sigma_8^2 = 8$, $\sigma_9^2 = 9$ and $\sigma_{10}^2 = 10$, with $\rho_1 = 0.8$, $\rho_2 = 0.725$, $\rho_3 = 0.65$, $\rho_4 = 0.575$, $\rho_5 = 0.5$, $\rho_6 = 0.425$, $\rho_7 = 0.35$, $\rho_8 = 0.275$ and $\rho_9 = 0.2$ to obtain

$$
\begin{bmatrix}
1 & & & & & & & & & \\
0.1314 & 2 & & & & & & & & \\
1.0046 & 1.7759 & 3 & & & & & & & \\
0.7540 & 1.3329 & 2.2517 & 4 & & & & & & \\
0.4847 & 1.8569 & 1.4475 & 2.5715 & 5 & & & & & \\
0.2655 & 0.4693 & 0.7928 & 1.4085 & 2.7386 & 6 & & & & \\
0.1219 & 0.2154 & 0.3640 & 0.6466 & 1.2572 & 2.7543 & 7 & & & \\
0.0456 & 0.0806 & 0.1362 & 0.2419 & 0.4704 & 1.0306 & 2.6912 & 8 & & \\
0.0133 & 0.0235 & 0.0397 & 0.0706 & 0.1372 & 0.3006 & 0.7640 & 2.3335 & 9 & \\
0.0028 & 0.0050 & 0.0084 & 0.0149 & 0.0289 & 0.0634 & 0.1611 & 0.4919 & 1.8974 & 10
\end{bmatrix}
$$

Acknowledgements

References

1. Skene SS, Kenward MG. The analysis of very small samples of repeated measurements. i. An adjusted sandwich estimator. *Statistics in Medicine*. DOI: 10.1002/sim.4073.

2. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**:983–997. DOI: 10.2307/2533558.

3. Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 2009; **53**:2583–2595. DOI: 10.1016/j.csda.2008.12.013.

4. SAS Institute Inc. 2004. *SAS/STAT® 9.1 User's Guide* Cary, NC: SAS Institute Inc.

5. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 2002; **21**:1429-1441. DOI: 10.1002/sim.1142.

6. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126-134. DOI: 10.1111/j.0006-341X.2001.00126.x.

7. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in one-way classification. *The Annals of Mathematical Statistics* 1954; **25**:290–302. DOI: 10.1214/aoms/1177728786.

8. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, ii. effects of inequality of variance and of correlation between errors in two-way classification. *The Annals of Mathematical Statistics* 1954; **25**:484–498. DOI: 10.1214/aoms/1177728717.

9. Bellavance F, Tardif S, Stephens MA. Tests for the analysis of variance of crossover designs with correlated errors. *Biometrics* 1996; **52**:607–612. DOI: 10.2307/2532899.

10. Chen, X, Wei L. A comparison of recent methods for the analysis of small sample cross-over studies. *Statistics in Medicine* 2003; **22**:2821–2833. DOI: 10.1002/sim.1537.

11. Scheffé H. A method for judging all contrasts in the analysis of variance. *Biometrika* 1953; **40**:87–104. DOI: 10.1093/biomet/40.1-2.87.

12. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer, 2000.

13. Satterthwaite FE. Synthesis of Variance. *Psychometrika* 1941; **6**:309–316. DOI: 10.1007/BF02288586.

14. Jones B, Kenward MG. *Design and analysis of cross-over trials* (2nd edn). Chapman and Hall/CRC, 2003.

15. Stuart A, Ord JK. *Kendall's advanced theory of statistics: vol 1. Distribution Theory* (6th edn). Hodder Arnold, 1994.

16. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis* Wiley, 2004.

17. Crowder MJ, Hand DJ. *Analysis of Repeated Measures* Chapman and Hall, 1990.

18. Nelder JA. The analysis of randomized experiments with orthogonal block structure. i. block structure and the null analysis of variance. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 1965; **283**:147–162. DOI: 10.1098/rspa.1965.0012.

19. Nelder JA. The analysis of randomized experiments with orthogonal block structure. ii. treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 1965; **283**:163–178. DOI: 10.1098/rspa.1965.0013.

20. Geisser S, Greenhouse SW. An extension of Box's results on the use of the F distribution in multivariate analysis *The Annals of Mathematical Statistics* 1958; **29**:885–891. DOI: 10.1214/aoms/1177706545.

21. Greenhouse SW, Geisser S. On methods in the analysis of profile data. *Psychometrika* 1959; **24**:95–112. DOI: 10.1007/BF02289823.

22. Huynh H, Feldt LS. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics* 1976; **1**:69–82. DOI: 10.3102/10769986001001069.

23. Brammer RJ. Modelling covariance structure in ascending dose studies of isolated tissues and organs. *Pharmaceutical Statistics* 2003; **2**:103–112. DOI: 10.1002/pst.30.