Biostatistics (2010), **11**, 1, *pp.* 1–17
doi:10.1093/biostatistics/kxp046
Advance Access publication on November 13, 2009

The use of baseline covariates in crossover studies

MICHAEL G. KENWARD*

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, University of London,
Keppel Street, London, WC1E 7HT, UK
mike.kenward@lshtm.ac.uk

JAMES H. ROGER

Research Statistics Unit, GlaxoSmithKline, Berkley Avenue, Greenford, Middlesex UB6 0NN, UK

Summary

It is our experience that in many settings, crossover trials that have within-period baseline measurements are analyzed wrongly. A “conventional” analysis of covariance in this setting uses each baseline as a covariate for the following outcome variable in the same period but not for any other outcome. If used with random subject effects such an analysis leads to biased treatment comparisons; this is an example of cross-level bias. Using a postulated covariance structure that reflects the symmetry of the crossover setting, we quantify such bias and, at the same time, investigate potential gains and losses in efficiency through the use of the baselines. We then describe alternative methods of analysis that avoid the cross-level bias. The development is illustrated throughout with 2 example trials, one balanced and orthogonal and one highly unbalanced and nonorthogonal.

Keywords: Analysis of covariance; Causal pathway; Covariance structure; Cross-level bias; Experimental design; REML; Restricted maximum likelihood.

1. Introduction and examples

Crossover trials have been widely used in medical research for many years for the assessment of reversible treatments for comparatively stable conditions. The literature on such designs is now large, with 2 standard texts on the subject, Senn (2002) and Jones and Kenward (2003). A crossover study is distinguished from the conventional parallel group design in having each subject randomly allocated to a “sequence” of treatments. In spite of this difference, however, the goal of a crossover study, in nearly all settings, remains the same as that of a parallel-group study to compare the effects of the “single” treatments, not the effects of the sequences to which the subjects are randomized. Thus, in a crossover trial, subjects are not randomized to the interventions under comparison and this has important implications for analysis and interpretation. In particular, the crossover design shares important features with observational studies, yet there is a tendency for those analyzing data from such studies to use modes of reasoning acquired from completely randomized designs. This can lead to misconceptions and errors in the analysis. In this paper, we focus on one such issue, the incorporation of baseline measurements into the analysis.

*To whom correspondence should be addressed.

Such covariates may be measured either at the beginning of the study (prior to randomization), which we term prerandomization, or at the beginning of each treatment period, which we term period-dependent. Commonly they will consist of the realizations of the outcome or response measurement itself, but as in a parallel-group study, measurements on other quantities may also be collected, and may or may not be period-dependent.

We will illustrate the results in the paper with 2 examples; the first, a simple balanced complete design and the second, a highly unbalanced design with missing data.

The first example is taken from Jones and Kenward (2003, Example 5.6), and is from a trial for the comparison of 3 treatments for high blood pressure, comprising the trial drug at 20 mg and 40 mg doses, respectively, and a placebo. Two complete replicates of the Williams arrangement (Jones and Kenward, 2003, Section 4.2.1) were used in which all 6 possible sequences occur, that is, with 3 periods and 12 subjects in total. The original authors give no information on the overall length of the washout or treatment periods. The measurements are of systolic blood pressure (in mm Hg), measured under each treatment at 10 successive times: 30 and 15 min before treatment, and at 8 times posttreatment. Here, we use the average of the 2 pretreatment measurements as the period-dependent baselines, and the measurement at 90 min as the response. For the analyses, we will focus on the high- versus low-dose treatment comparison.

The second example is taken from a first-in-human single-dose study organized as a 5-period crossover in 2 cohorts, with 10 healthy volunteers in each cohort. This data set appeared in a plenary data analysis session at the 2006 conference of Statisticians in the Pharmaceutical Industry (PSI) and has been made available as supplementary material at *Biostatistics* online. The trial had 8 treatments: a negative control (P), a positive control (G), and 6 increasing doses (A, B, C, D, E, and F), 3 in each cohort organized as ascending doses within each subject, alternating between cohorts. The doses are 10, 30, 60, 150, 250, and 400 mcg of an inhaled substance. The design has 2 replicates of 5 sequences within each cohort of 10 subjects. Three subjects, on the sequences ACPEG, GBDFP, and BDPFG, respectively, withdrew after 1 period and 3 replacement subjects were added using the same 3 sequences. These subjects completed all 5 periods. One subject, on GBDFP, withdrew after 3 periods and was not replaced. Doses A, C, and E only appear in Cohort 1, while doses B, D, and F only appear in Cohort 2. Both the negative and positive control appear in every sequence. The design is as follows:

Cohort 1

Period	1	2	3	4	5
N = 2	P	A	G	C	E
N = 2	G	A	C	E	P
N = 2+	A	C	P	E	G
N = 2	A	P	C	G	E
N = 2	A	G	C	P	E

Cohort 2

Period	1	2	3	4	5
N = 2	P	B	G	D	F
N = 1++	G	B	D	F	P
N = 2+	B	D	P	F	G
N = 2	B	P	D	G	F
N = 2	B	G	D	P	F

There was a 1 week interval between each test day, meaning that subjects in each cohort had a 14-day washout interval. Within each period, lung function was measured using forced expiratory volume in 1 s (FEV_1 litres) 7 times, at baseline (0), 2, 6, 9, 12, 22, and 24 h. All subjects were dosed at time zero, just after their baseline measurement had been taken. This start time was standardized across periods within subject. The actual start time varied by up to an hour from subject to subject. Here, as the outcome variable, we only consider the FEV_1 measurement taken at 12 h. The design is highly unbalanced, so we will focus on 2 treatment contrasts that reflect very different sources of information: (i) the difference between the positive and negative controls, both of which all subjects received (except those who withdrew), with a design efficiency of 97% and (ii) the difference between the high and low doses, for which no subject received both, with a design efficiency of 46%. We expect within-subject information to dominate in estimating the former, and between-subject information to be relatively more important for the latter.

In Section 2, we set out a general modeling framework for the crossover setting with baselines. In particular, we formulate an appropriate covariance structure for such data and fit the structure to a range of examples, including the 2 illustrative examples introduced above. In Section 3, we briefly review the use of baseline covariates in parallel-group studies, and we see how these ideas carry over to the particular case of prerandomization covariates in crossover studies. In Section 4, we then explore the role of period-dependent covariates, in terms of change-from-baseline analyses, their use as covariates (conditional) and as outcomes (unconditional). We also touch on the relevance of missing data from a modeling (not inferential) perspective. We close in Section 5 with a brief discussion and summary of the main conclusions.

2. Notation and basic results

Suppose that we have a crossover trial with p periods. Denote by $\mathbf{Y}_{ik}$, the vector of p response measurements from the periods under active treatment from the kth subject in sequence $i, i = 1, \ldots, s$. We assume that both period and direct treatment effects are included in the model, but no other effects, such as carry over. Similarly, denote by $\mathbf{X}_{ik}$, the corresponding set of baseline measurements. Denote the individual measurements similarly by X_{ijk} and Y_{ijk} for $j = 1, \ldots, p$. We assume in the following that these have a joint multivariate Gaussian distribution although, strictly, when using analysis of covariance, we require only that the Ys have the appropriate conditional joint Gaussian distribution given the Xs.

We further assume that $\mathrm{E}(\mathbf{Y}_{ik}) = \mathbf{A}_i\boldsymbol{\beta}$ for design matrix $\mathbf{A}_i$ associated with the ith sequence and $\boldsymbol{\beta}$, the corresponding parameters, and $\mathrm{E}(\mathbf{X}_{ik}) = \boldsymbol{\phi}$, a $(p \times 1)$ vector of means corresponding to the p times of the baseline measurements. Then,

$$\begin{pmatrix} \mathbf{X}_{ik} \\ \mathbf{Y}_{ik} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \boldsymbol{\phi} \\ \mathbf{A}_i\boldsymbol{\beta} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^T & \boldsymbol{\Sigma}_{YY} \end{pmatrix} \right\}. \tag{2.1}$$

Three important expressions immediately follow from this. First, the marginal distribution of the Ys alone:

$$\mathbf{Y}_{ik} \sim \mathrm{N}\left(\mathbf{A}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_{YY}\right). \tag{2.2}$$

Second, the marginal distribution of the p changes from baseline:

$$\mathbf{Y}_{ik} - \mathbf{X}_{ik} \sim \mathrm{N}(\mathbf{A}_i\boldsymbol{\beta} - \boldsymbol{\phi}; \boldsymbol{\Sigma}_{YY} + \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{XY}^T). \tag{2.3}$$

Third, the conditional distribution of $\mathbf{Y}_{ik}$ given $\mathbf{X}_{ik}$:

$$\mathbf{Y}_{ik} \mid \mathbf{X}_{ik} \sim \mathrm{N}\{\mathbf{A}_i\boldsymbol{\beta} - \boldsymbol{\theta}(\boldsymbol{\phi} - \mathbf{X}_{ik}); \quad \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{XY}^T\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\}, \tag{2.4}$$

for $\boldsymbol{\theta} = \boldsymbol{\Sigma}_{XY}^T\boldsymbol{\Sigma}_{XX}^{-1}$. The conditional distribution of the differences $\mathbf{Y}_{ik} - \mathbf{X}_{ik}$ given the baselines $\mathbf{X}_{ik}$ is identical to (2.4), except that the mean is shifted by $\mathbf{X}_{ik}$.

The comparative behavior of the 3 analyses that can be derived from these representations: (1) analysis of the outcome Y only, (2) analysis of change from baseline $Y - X$, and (3) analysis of Y conditional on X, depends on the model chosen and on the size and form of the covariance matrices $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$, and the covariances in $\boldsymbol{\Sigma}_{XY}$. We explore the relevance of particular forms for these in the following. In particular, we begin by making the assumption of variance and covariance homogeneity within the Xs and within the Ys, and between the Xs and the Ys. In other words, we assume a uniform, or compound symmetry, structure for each of 3 matrices:

$$\begin{aligned}\boldsymbol{\Sigma}_{XX} &= \sigma_{xx}\mathbf{I}_p + \eta_{xx}\mathbf{J}_p,\\ \boldsymbol{\Sigma}_{YY} &= \sigma_{yy}\mathbf{I}_p + \eta_{yy}\mathbf{J}_p,\\ \boldsymbol{\Sigma}_{XY} &= \sigma_{xy}\mathbf{I}_p + \eta_{xy}\mathbf{J}_p = \boldsymbol{\Sigma}_{XY}^T,\end{aligned} \tag{2.5}$$

where $\mathbf{I}_p$ and $\mathbf{J}_p$ denote the $p \times p$ identity matrix and matrix of ones, respectively. We regard these as the least constrained forms for these matrices that are compatible with the stability assumption inherent in most analyses for crossover trials. Note that these expressions do not constrain the baseline and response variables to have the same variances, although they are each assumed constant across periods. Nor is the correlation between an adjacent baseline and a response measurement assumed to be the same as that between a baseline and a response from a different period. To see this, more precisely, consider the following variances and covariances implied by (2.5):

$$\mathrm{V}(Y_{ijk}) = \sigma_{yy} + \eta_{yy}, \quad \mathrm{V}(X_{ijk}) = \sigma_{xx} + \eta_{xx},$$

$$\mathrm{Cov}(Y_{ijk}, X_{ijk}) = \sigma_{xy} + \eta_{xy}, \quad \mathrm{Cov}(Y_{ijk}, X_{ij'k}) = \eta_{xy}.$$

From this, it can be seen that σ_{xy} determines the additional covariance due to a baseline and response being from the same period; we will call these the “associated” baseline and response measurements, indicating that they come from the same period.

We can get additional insight into this structure through the reformulation of the joint model for the Ys and Xs in (2.1), imposing the structure in (2.5) on $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{XY}$, and $\boldsymbol{\Sigma}_{YY}$. This is conveniently done in terms of two 2×2 unstructured covariance matrices, $\boldsymbol{\Sigma}_{\mathrm{W}}$ and $\boldsymbol{\Sigma}_{\mathrm{B}}$, where

$$\boldsymbol{\Sigma}_{\mathrm{W}} = \begin{pmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{xy} & \sigma_{yy}\end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathrm{B}} = \begin{pmatrix}\eta_{xx} & \eta_{xy}\\ \eta_{xy} & \eta_{yy}\end{pmatrix}. \tag{2.6}$$

If the observations from a subject, $\mathbf{Z}_{ik}$ say, are then ordered by time, that is, $\mathbf{Z}_{ik} = (X_{ik1}, Y_{ik1}, X_{ik2}, \dots, X_{ikp}, Y_{ikp})^T$, then in terms of these 2 covariance matrices

$$\mathrm{V}(\mathbf{Z}_{ik}) = \boldsymbol{\Sigma}_{\mathrm{W}} \otimes \mathbf{I}_p + \boldsymbol{\Sigma}_{\mathrm{B}} \otimes \mathbf{J}_p. \tag{2.7}$$

In this way, $\boldsymbol{\Sigma}_{\mathrm{W}}$ represents the covariance matrix for X and Y within a full treatment period and $\boldsymbol{\Sigma}_{\mathrm{B}}$, the covariance matrix between baseline and response at the subject level. To fit the model in (2.1) with this covariance structure, a fixed-effects model is then constructed with a categorical time effect with $2p$ levels, 1 for each element of $\mathbf{Z}$, and a categorical treatment term which applies to the Ys only. We give an example using the SAS MIXED procedure for fitting such a model in the Appendix.

Many of our conclusions below will depend on the actual values taken by the 6 covariance parameters in (2.5). It is therefore of interest to view estimates of these taken from a range of examples, with very different types of endpoint. One thing to note is that the design of all these studies is a crossover, so there is an automatic bias toward those where we may expect a strong subject effect. We present such estimates in Table 1, which have been estimated from the 2 illustrative examples introduced above, and from a further 3 trials, 3–5, using very different types of measurement; FEV$_1$ again and 3 new endpoints: Nitric oxide, a pain visual analogue score (VAS), and QTc. The last of these is the measured time between the start of the Q wave and the end of the T wave in the heart’s electrical cycle, corrected for the heart rate.

Where repeated measurements have been collected within periods, we have presented the covariance parameter estimates separately for each repeated measurement; these will of course share the same baseline measurement. These illustrate how correlations change across time, usually decaying as the observation moves further away from the baseline measurement.

Table 1. *Covariance parameters and correlations estimated from* 5 *example crossover trials*

Source	Covariance parameter						XY correlation	
	$\hat{\sigma}_{xx}$	$\hat{\eta}_{xx}$	$\hat{\sigma}_{yy}$	$\hat{\eta}_{yy}$	$\hat{\sigma}_{xy}$	$\hat{\eta}_{xy}$	W-P	B-P
Example 1: Systolic blood pressure ($N = 12$)	51.79	59.31	55.77	101.45	−12.14	64.06	−0.2260	0.8259
Example 2: FEV$_1$ ($N = 21$)	0.0276	0.4534	0.0233	0.4992	0.0109	0.4756	0.4298	0.9997
Example 3: Nitric Oxide ($N = 18$)								
Day 1 hour 1	0.0528	0.5348	0.0511	0.4774	0.0495	0.5046	0.9524	0.9986
Day 1 hour 2	0.0528	0.5348	0.0452	0.4608	0.0464	0.4943	0.9511	0.9957
Day 7 hour 1	0.0528	0.5348	0.0368	0.4545	0.0051	0.4966	0.1163	1.0072
day 7 hour 2	0.0528	0.5348	0.0335	0.4607	0.0051	0.5013	0.1214	1.0100
Example 3: FEV$_1$ ($N = 21$)								
Day 7 hour 0	0.0629	0.4954	0.0192	0.6065	0.0071	0.5393	0.2050	0.9840
Day 7 hour 1	0.0629	0.4954	0.0164	0.6331	0.0036	0.5441	0.1119	0.9716
Day 7 hour 2	0.0629	0.4954	0.0411	0.6458	0.0058	0.5539	0.1133	0.9793
Example 4: VAS ($N = 43$)								
Week 1	0.6825	1.0179	1.1392	1.3001	0.4055	0.9059	0.4599	0.8397
Week 2	0.6825	1.0078	1.6368	1.5112	0.1047	1.2959	0.0991	1.0449
Example 5: QTcF ($N_1 = 145$, $N_2 = 232$)								
Set 1a hour 0.5	64.803	211.48	61.179	201.39	4.175	203.52	0.0663	0.9862
Set 1a hour 1	52.755	211.34	58.293	207.79	10.769	199.78	0.1942	0.9534
Set 1a hour 2	60.924	213.42	58.998	216.38	1.709	212.15	0.0285	0.9872
Set 1a hour 3	50.603	224.72	58.333	205.66	2.737	205.38	0.0504	0.9554
Set 1a hour 4	60.561	200.03	53.182	196.86	−6.187	198.82	−0.1090	1.0019
Set 1a hour 6	63.166	186.12	48.999	188.87	−0.780	188.17	−0.0140	1.0036
Set 1a hour 12	42.786	221.15	58.496	200.96	5.172	208.21	0.1034	0.9877
Set 1a hour 18	70.859	178.07	61.655	198.30	12.436	183.09	0.1882	0.9743
Set 1a hour 24	61.841	222.32	43.693	235.02	15.407	223.26	0.2964	0.9767
Set 1b hour 0.5	26.012	197.78	61.997	201.34	4.837	198.71	0.1205	0.9957
Set 1b hour 1	25.854	199.00	58.236	209.10	−0.095	203.48	−0.0024	0.9975
Set 1b hour 2	25.626	200.00	59.558	215.48	0.832	205.59	0.0213	0.9903
Set 1b hour 3	26.015	198.21	58.190	205.85	1.371	200.38	0.0352	0.9920
Set 1b hour 4	25.992	198.39	53.955	196.10	0.963	194.30	0.0257	0.9851
Set 1b hour 6	25.485	198.00	49.697	189.97	0.927	183.59	0.0260	0.9466
Set 1b hour 12	26.475	234.58	58.037	201.66	5.048	204.99	0.1288	0.9425
Set 1b hour 18	23.856	181.74	61.991	197.08	7.653	180.66	0.1990	0.9546
Set 1b hour 24	27.209	232.57	44.180	234.40	7.231	218.33	0.2086	0.9351
Set 2 hour 0.5	46.420	182.62	72.845	182.67	20.131	182.32	0.3462	0.9982
Set 2 hour 1	46.373	182.75	64.495	182.77	22.229	179.89	0.4065	0.9843
Set 2 hour 2	46.248	182.82	50.689	183.94	17.285	181.87	0.3570	0.9918
Set 2 hour 3	46.446	182.53	57.164	168.22	17.708	173.42	0.3437	0.9897
Set 2 hour 6	45.612	182.78	53.149	191.74	6.282	168.44	0.1276	0.8997
Set 2 hour 24	46.334	183.36	51.044	179.77	2.817	176.61	0.0579	0.9727

W-P: within-period, B-P: between-period.

2.1 *Example 3*

This is a single center, randomized, double-blind, placebo-controlled, and 2 × 2 crossover study (AB/BA) to investigate the effect of treatment with repeat doses of a treatment on bronchial hyperreactivity to adenosine monophosphate (AMP) challenge. The AMP challenges were carried out 2 h after dosing on Day 7 and again 24 h after this final dosing. All measurements presented here were obtained prior to the AMP challenges. Each period consists of 7 days of dosing, with baseline measurements made in each period prior to dosing on Day 1. Measurements are made after dosing on Day 1 and also on Day 7, and there is a 14-day washout period between periods. We consider here 2 secondary endpoints: (1) the exhaled nitric oxide (log of the arithmetic mean of 3 observations) at 1 and 2 h on both Day 1 and Day 7 and (2) FEV$_1$ (litres) measured at 0 (prior to Day 7 dose), 1, and 2 h on Day 7. To maintain common baseline variances, all subjects who do not have a complete set of data for all repeated measurements, separately for each endpoint, have been removed.

2.2 *Example 4*

This is a 2 × 2 crossover study on the treatment of pain. The treatment period consisted of 2 weeks with a washout period of 2 to 4 weeks. The outcomes are the weekly average of a daily VAS, for weeks 1 and 2. Again only completers were included.

2.3 *Example 5*

This is based on six 2 × 2 crossover trials, all comparing placebo to a single dose of 400 mg moxifloxacin, which is a positive control for QT elongation. Washout between the 2 periods was 14 days in 2 cases, 7 days in 3 cases, while the remaining trial specified a washout of 3 to 7 days. They are taken from a meta-analysis of 9 trials (Stylianou *and others*, 2008). The 6 trials separate into 2 sets of 3. The first 3 have a 24 h baseline where the same sequence of measurements is made 24 h earlier. So each repeated measurement has a matching predose baseline repeated measurement. These have been analyzed using as baseline the individual time matched baselines (set 1a), and also using throughout the average of the predose measurements from the previous 24 hours (set 1b). In the second set, there is a pair of predose observations, the average of which is used as the baseline. For simplicity, the gender effect has been removed from the models used in Stylianou *and others* (2008).

As well as the estimates of the 6 covariance parameters, the associated within-period and between-period correlations ($\rho_{\mathrm{W}} = \sigma_{xy}/\sqrt{\sigma_{xx}\sigma_{yy}}$ and $\rho_{\mathrm{B}} = \eta_{xy}/\sqrt{\eta_{xx}\eta_{yy}}$) are presented in Table 1. Note that some correlation estimates are greater than one: the method of estimation does not constrain these to the $(-1, 1)$ interval; these should be interpreted as values very close to one.

Several broad points can be made from Table 1. The subject-level correlations (ρ_{B}) are consistently very high indeed, often approaching one. Also, they do not vary across time for the 4 endpoints measured repeatedly. By contrast, the within-period correlations are very much smaller, in some cases negligible. This represents dependency that exists locally in time, once the overall subject-level dependency has been accounted for. These correlations tend to decay with time as we move away from baseline measurement. But also the correlation roughly a day later (18 or 24 h) can sometimes increase again (e.g. QTc in Example 5). This commonly observed feature might be due to some personalized diurnal rhythm. The variability of the Xs and the Ys within a trial can be very similar, or quite different, depending on the example. We return to a detailed discussion of the values from Examples 1 and 2 when we consider the analysis of these trials in more detail below.

To simplify the essential discussion, we focus on crossover designs that are variance balanced with respect to treatment, that is, for which all normalized treatment contrasts have the same variance. This means that in considering the precision of different approaches to estimation of treatment effects, we need to consider only a single treatment contrast, τ say, whose exact form is irrelevant to the comparisons of efficiency. Depending on the particular design used, and in the absence of period-dependent covariates, information on τ may be present only in the within-subject stratum or in both the within-subject and between-subject strata. We refer to designs that lead to the former as orthogonal and to the latter as nonorthogonal, noting that, more generally, these terms do have a wider use and meaning. The first illustrative example is both variance balanced and orthogonal, the second neither of these.

We assume that τ is estimated using generalized least squares (GLS) with the appropriate known covariance matrix. In practice, a consistent estimator of this covariance matrix would be used, implying that the given measures of precision only hold asymptotically. This is acceptable for our goal of making comparisons of precision of different estimators. When these methods are applied in practice, however, some small sample adjustment will be needed in some settings for the estimated variances and associated inferences; the appropriate corrections are given in Kenward and Roger (1997, 2009).

3. Prerandomization covariates

In the context of parallel-group studies, 2 main roles for adjustment for baseline covariates have been widely discussed, for example, Pocock *and others* (2002); namely reducing variability and adjusting for baseline imbalance. In the current crossover context of prerandomization covariates, we are concerned only with the former, that is, we assume that the role is to explain variability in the outcome measurements and so increase precision in the treatment estimators. In the crossover setting, a prerandomization covariate can influence only between-subject information and so has strictly limited value. This is in contrast to the situation we meet below with period level baselines, which can influence both between- and within-subject information. In an orthogonal design, such as a complete Williams square, all information on the treatment effects is wholly within-subject, and so a prerandomization covariate will have no effect at all on the treatment estimators and their precision. It is surprising that this very obvious property of the orthogonal crossover design is still not more widely understood. For example, in a very recent and influential publication, McCann *and others* (2007), pointless adjustment is made for several prerandomization covariates in the analysis of data from a Williams design with, predictably, a negligible effect on the results. In this example, like many, a few missing observations remove the strict orthogonality of the design, but the information that is then present in the between-subjects stratum is still nugatory, and its existence does not change the overall message. When analyses need to be prespecified, however, it is possible that such baselines might be included as covariates to allow for the possibility of extensive dropout.

Adjustment for prerandomization covariates will therefore make a worthwhile improvement to precision when there exists a nontrivial amount of information on the treatment effect in the between-subjects stratum, which requires a highly nonorthogonal design, such as so-called “incomplete” designs for which $p < t$ (see Jones and Kenward, 2003, Section 4.2.2), or when there is a very high proportion of dropouts from an originally orthogonal one. Consider, as an illustration, the balanced 3-treatment 2-period design with each replicate consisting of all 6 possible sequences. With a single replicate of this design, and using the model introduced in Section 2, the GLS estimator of any treatment difference has variance

$$\frac{\sigma_{yy}(\sigma_{yy} + 2\eta_{yy})}{2\sigma_{yy} + 3\eta_{yy}} = \sigma_{yy}\left(\frac{1+\rho_Y}{2+\rho_Y}\right) \tag{3.1}$$

for $\rho_Y = \eta_{yy}/(\sigma_{yy} + \eta_{yy})$. Even in the impossibly extreme case in which the covariate removes the between-subject variability altogether, the ratio of variances of the unadjusted and adjusted estimators is still only

$$1 + \frac{\rho_Y}{2+\rho_Y}$$

which, with a correlation of 0.8, for example, is less than 1.3. From the estimates in Table 1, this quantity is remarkably stable, ranging from 1.19 to 1.33. In practice, adjustment will be less effective than this, possibly much less so. Thus, we do not usually expect a large gain in precision in such settings, but adjustment may still be considered worthwhile in some examples. The problem is that the variability that is being explained, that is, the between-subject component, is usually the far less important part of the information contributing to the treatment estimator, even in nonorthogonal designs. On the other hand, such adjustment should have a negligible effect as well on power, so it can be regarded as a relatively safe, if pointless, exercise.

4. Period-dependent baseline covariates

We now consider the use of period-level baselines. In contrast to Section 3, we are now considering the impact of these with respect to information at the between-subject and within-subject level, hence, both variance reduction and baseline imbalance between treatments are potentially important.

We also need at this point to clarify the use of baseline as a term. It is sometimes used to refer to a measurement made in a trial (parallel or crossover) as part of the measurement process but following treatment. Examples are measurements made at the start of exercise tests and before challenges. Such baselines should never be used as covariates, and we exclude them from the current development. Although the period-dependent baselines we consider here are made following randomization of the subject to sequence, we are assuming that they are made before treatment is given within each period and that the washout periods are sufficient to ensure that these baselines are not influenced by previous treatment, that is, they are not affected by carryover. There are situations in which the washout is insufficiently long firmly to rule out the possibility of baseline contamination through carryover effects. Such contamination obviously makes such baselines unsuitable as covariates. Although one might test for this formally before deciding whether even to consider incorporating the baseline measurements in the analysis, such a sequential procedure raises many of the issues associated with prior testing in the 2-period 2-treatment design (Senn, 1988; Freeman, 1989) and should therefore be avoided.

4.1 *Change from baseline*

We begin by considering a conventional analysis of the change from baseline, $D = Y - X$. It follows immediately from (2.3) that any least squares estimator of the direct treatment effect will be unbiased, as the impact of $\boldsymbol{\phi}$ in the expectation will be only to modify the intercept and period effects. Hence, the design matrices for both approaches here coincide. The consequence on the analysis, in comparison with that of the Ys only, will therefore only be in terms of precision. Under the covariance assumptions in (2.5), we see that the p changes $\mathbf{D}_{ik}$ for one subject have the compound symmetry covariance structure:

$$\mathrm{V}(\mathbf{D}_{ik}) = \mathrm{V}(\mathbf{Y}_{ik} - \mathbf{X}_{ik}) = (\sigma_{xx} + \sigma_{yy} - 2\sigma_{xy})\mathbf{I}_p + (\eta_{xx} + \eta_{yy} - 2\eta_{xy})\mathbf{J}_p,$$

compared with that of the original Ys:

$$\mathrm{V}(\mathbf{Y}_{ik}) = \sigma_{yy}\mathbf{I}_p + \eta_{yy}\mathbf{J}_p.$$

It follows that in any given setting, depending on the relative sizes of these parameters, and on the particular design, either an analysis of the Ys or of the changes from baseline $Y - X$, may lead to more precise estimators of the treatment effects. We need to consider what is likely to occur in common settings. Suppose first that the design is orthogonal in the sense introduced in Section 2. In such designs, the treatment estimators do not depend on the variance parameters. Let $\widehat{\tau}_Y$ and $\widehat{\tau}_D$ denote, respectively, the GLS estimators from the response data alone and the changes from baseline. It follows immediately that

$$R = \frac{\mathrm{V}(\widehat{\tau}_D)}{\mathrm{V}(\widehat{\tau}_Y)} = \frac{\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy}}{\sigma_{yy}}.$$

The analysis of the changes D will be more precise only when $\sigma_{xx} < 2\sigma_{xy}$. Recall that σ_{xx} is the within-subject, or residual, variance of the baselines, and σ_{xy} represents the amount by which the covariance of an associated baseline and response exceeds that of a response and baseline from different periods. This inequality holds for only 3 of the examples in Table 1. These are the 2 Day 1 measurements of nitric oxide in Example 3, and the Week 1 VAS in Example 4. In both cases, this feature has decayed by later time points. Both parameters relate to the within-subject distribution and the property is generally about the size of the local rather than long-term (or between-subject) correlation. If the variance of baseline is the same as that of the response, then this is equivalent to the within-subject correlation of response with baseline being greater than a half.

Suppose that we have a completely uniform covariance structure across all 8 measurements, as would be implied by a simple random subject effects model. This might be appropriate when the time interval between baseline and associated response is of the same order as between a response and a baseline from the following period, or both long enough although of different lengths, and the variance is stable across all measurements. This would imply that $\sigma_{yy} = \sigma_{xx}$, $\eta_{yy} = \eta_{xx} = \eta_{xy}$, and $\sigma_{xy} = 0$. We call this the "uniform assumption." Under this, we have the ratio

$$R = \frac{\mathrm{V}(\widehat{\tau}_D)}{\mathrm{V}(\widehat{\tau}_Y)} = \frac{\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy}}{\sigma_{yy}} = \frac{\sigma_{yy} + \sigma_{yy} - 2 \times 0}{\sigma_{yy}} = 2;$$

the estimator from the differences D would have twice the variance of that based on the simple responses Y alone.

In conclusion, the analysis of the differences will only be worth considering in those settings in which $\sigma_{xx} < 2\sigma_{xy}$. This is most likely to happen if baselines are relatively close in time to that of the associated responses, compared to the gap between treatment periods, implying that σ_{xy} may plausibly be nonnegligible. Note that this does not apply in Example 5, with the QTc data, for which the within-subject correlation with 24-h previous baseline (set 1) is low. However, using same-day-averaged baseline, the within-subject correlation is higher (set 2). This highlights the importance of the local (or serial) nature of the correlation, for an analysis of difference from baseline to be appropriate. In addition to this, the inequality is also more likely to hold if the baseline variables are considerably less variable than the response measurements. This can happen when there is considerable heterogeneity of response to treatment among a group of carefully selected subjects or the baseline measurement has been averaged over several baseline measurements, as in Example 5 set 1b and set 2. Empirical examples of the former are given in Kenward and Jones (1987). Although the results given here depend on exact balance and orthogonality, mild departures from this due to dropout and incomplete replication of sequences would not be expected to change this overall picture.

As an illustration, we consider the first example, the 3-treatment Williams design which is balanced and orthogonal. The 6 covariance parameter estimates are presented in Table 1. Note that $\hat{\sigma}_{xy}$, the additional within-subject covariance for associated X and Y, is negative but comparatively small. Consequently, we would expect the analysis of change from baseline ($D = Y - X$) to be far less precise that the analysis of Y alone, and this is indeed what is found. Using restricted maximum likelihood (REML) to fit a random subject effects model with categorical period and direct treatment fixed effects, the high–low dose comparison is estimated (with accompanying standard error [SE]) for Y and D, as 6.67 (3.04) and 5.08 (4.61), respectively, showing a 52% increase in SE in the latter. Using Y alone, the effect is statistically significant at 5% ($P = 0.04$), while using D, it is far from significance ($P = 0.28$). The use of change from baseline is, for this example, clearly counterproductive.

When designs are used in which treatments and periods are highly nonorthogonal, such as Example 2 and those considered in Section 4, then it is harder to provide broad guidelines. In such settings, the GLS estimators consist of weighted combinations of within-subject and between-subject information, with the weights depending directly on the parameters of the covariance structure. It is possible that the changes from baseline may remove considerable variability from the subject sums component, and hence markedly increase the contribution of the between-subject information to the overall precision. However, this may well be counterbalanced by the potential loss in precision of the within-subject differences, as seen above in orthogonal designs. How these 2 contributions balance out in practice depends both on the covariance parameters and the particular design. Consider again, as an example, the balanced 3-treatment 2-period design with each replicate consisting of all 6 possible sequences. From (3.1), it can be seen that the ratio of the variances from the estimators using the differences $Y - D$ and Y alone, respectively, is in terms of the original covariance parameters,

$$R = \frac{(2\sigma_{yy} + 3\eta_{yy})(\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy})(\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy} + 2\eta_{xx} + 2\eta_{yy} - 4\eta_{xy})}{\sigma_{yy}(\sigma_{yy} + 2\eta_{yy})(2\sigma_{yy} + 2\sigma_{xx} - 4\sigma_{xy} + 3\eta_{xx} + 3\eta_{yy} - 6\eta_{xy})}$$
$$= \frac{(\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy})}{\sigma_{yy}} \frac{(1+\rho_D)}{(2+\rho_D)} \frac{(2+\rho_Y)}{(1+\rho_Y)}, \tag{4.1}$$

where

$$\rho_D = \frac{\eta_{yy} + \eta_{xx} - 2\eta_{xy}}{\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy} + \eta_{yy} + \eta_{xx} - 2\eta_{xy}}$$

is the within-subject correlation of the differences $Y - X$. Suppose now that the uniform assumption holds, that is, as before, $\sigma_{yy} = \sigma_{xx}$, $\eta_{yy} = \eta_{xx} = \eta_{xy}$, and $\sigma_{xy} = 0$. Under these assumptions, R in (4.1) is equal to $(2 + \rho_Y)/(1 + \rho_Y)$ which, for $0 < \rho_Y < 1$, is in the range $1.5 < R < 2$. Again, we see that the use of the changes from baseline increases the variance compared with the use of the Ys alone. The reduction in between-subject variability makes some contribution, and so R is smaller than the value 2 seen above for the orthogonal crossover designs, but never enough to counterbalance the increase in within-subject variability. This is an inefficient design, with considerable between-subject information. Even in this setting, we see that the use of change from baseline, under the uniform assumption, increases the variability. Given this, we conjecture that only with the most extreme inefficient designs would the use of change from baseline improve the precision relative to the use of Y alone, and our recommendation, if the uniform assumption approximately holds, is to use the latter not the former even with nonorthogonal designs.

In the general covariance setting (2.5), it is harder to draw broad conclusions because of the relative complexity of (4.1). However, if we again assume that $\sigma_{xx} < 2\sigma_{xy}$, the condition under which the differences lead to more precise estimates than Y alone for orthogonal designs, then a sufficient condition for greater precision from the analysis of the differences is that $\rho_D < \rho_Y$, which in turn implies

$$\left(\frac{1-\rho_Y}{\rho_Y}\right)(\eta_{xx} - 2\eta_{xy}) < \sigma_{xx} - 2\sigma_{xy} < 0.$$

The first term on the left-hand side, $(1 - \rho_Y)/\rho_Y$ will always be less than 1 for $\rho_Y > 0.5$, and this implies that, depending on the size of ρ_Y and η_{xy}, the covariance between any nonassociated X and Y must not be much smaller than η_{xx}, the covariance between any 2 Xs. Given that an X must be measured in the interval between any other X and a nonassociated Y, and given the symmetry in the assumed covariance structure, it is plausible that η_{xy} will typically be less than η_{xx}, so this condition is not one which might automatically be assumed to hold.

We now consider the second example, on lung function, in the light of these results. A random subject effects model with categorical period and direct treatment fixed effects has again been fitted using REML. We note first that the estimated residual and between-subject variance estimates are 0.023 and 0.503 for the raw outcomes Y, and 0.030 and 0.001 for the differences D, respectively. There is very high correlation between the baselines X and outcomes Y, and virtually all random between-subject variability has been removed from the differences. The covariance parameter estimates are given in Table 1. Note, the size of $\hat{\sigma}_{xy}$: this strongly suggests that the uniform assumption does not hold here. The key quantity $\sigma_{xx} - 2\sigma_{xy}$ is here estimated as 0.0057 which would suggest, for an orthogonal design, that subtracting the baseline would have very little effect on precision. Here, we would expect this to apply to the first (efficiently estimated) contrast. For the second contrast, it is harder to predict the effect of subtracting X. From the analyses of Y and D, respectively, we get, for contrast (1), the estimates (and SEs) 0.126 (0.049) and 0.224 (0.055). Although these estimates are rather different, the use of change from baseline has increased the variance only very slightly. For the second contrast, the estimates from Y and D are 0.055 (0.100) and

−0.014(0.099), respectively. Again the estimates are somewhat different but there is very little difference in precision. The differences we see between the treatment estimates (with and without change from baseline) represent the differences between the baseline scores for these contrasts (−0.081 and 0.072). In conclusion, analyzing change from baseline has had little effect on precision compared with the analysis of response alone. The consequences for the estimates are not negligible however, and this is probably due to the exceptionally high correlation between baseline and response at the subject level. We see in Section 4.3 that this is estimated as 0.997. This very high value has other implications for the analysis of these data which we will explore below.

The appropriate route, therefore, is to introduce the baselines as covariates, whether for the raw responses (Y) or the differences ($Y - X$). In this way any baseline imbalance is accommodated, but the data are being allowed to select between absolute and difference from baseline. This, however, introduces new issues that need to be addressed, as we see below.

4.2 *Baselines as covariates*

Without particular constraints on the covariance structure, all p observed baselines appear in the expectations of each element of $\mathbf{Y}_{ik}$ in the conditional distribution of $\mathbf{Y}_{ik}$ given $\mathbf{X}_{ik}$. This implies an analysis of covariance in which all p baselines appear as covariates for all p response variables. Such an analysis is rarely, if ever, done in practice however. Conventionally, only the associated baseline is used to adjust the corresponding response. We therefore begin, for the covariance structure introduced in Section 2, by considering what constraints on this structure would lead to this conventional analysis. From (2.4), we see that the regression coefficients of $\mathbf{Y}_{ik}$ on $\mathbf{X}_{ik}$ are given by the elements of $\boldsymbol{\theta} = \boldsymbol{\Sigma}_{XY}^T \boldsymbol{\Sigma}_{XX}^{-1}$, or in terms of the expressions for the covariance components in (2.5)

$$\begin{aligned}\boldsymbol{\theta} &= (\sigma_{xy}\mathbf{I}_p + \eta_{xy}\mathbf{J}_p)(\sigma_{xx}\mathbf{I}_p + \eta_{xx}\mathbf{J}_p)^{-1} \qquad (4.2)\\ &= \frac{\sigma_{xy}}{\sigma_{xx}}\mathbf{I}_p + \frac{\eta_{xy}\sigma_{xx} - \eta_{xx}\sigma_{xy}}{\sigma_{xx}(\sigma_{xx} + p\eta_{xx})}\mathbf{J}_p.\end{aligned}$$

For the conventional analysis of covariance model to hold, $\boldsymbol{\theta}$ must be diagonal because each element of $\mathbf{Y}_{ik}$ is regressed only on the element of $\mathbf{X}_{ik}$ from the same period, which in turn implies that $\eta_{xy}\sigma_{xx} - \eta_{xx}\sigma_{xy} = 0$. First, we note that this cannot hold under the uniform assumption, for which $\sigma_{xy} = 0$, unless all measurements are mutually independent. More generally, this requirement implies that $\boldsymbol{\Sigma}_{XY}$ is proportional to $\boldsymbol{\Sigma}_{XX}$, and it is difficult to find a practical justification for this rather contrived assumption.

We consider next, the behavior of the conventional analysis of covariance under the covariance structure of Section 2, (2.5). To explore potential bias, we again examine separately the within-subject and between-subject information. Let the fixed matrix $\mathbf{K}$ be any $p \times (p-1)$ matrix satisfying $\mathbf{K}^T\mathbf{K} = \mathbf{I}_{p-1}$ and $\mathbf{K}^T\mathbf{j}_p = \mathbf{0}$, for $\mathbf{j}_p$ a p-dimensional vector of one's. The within-subject information for the ikth subject can be represented by $\mathbf{K}^T\mathbf{Y}_{ik}$, which has regression model

$$\mathrm{E}(\mathbf{K}^T\mathbf{Y}_{ik} \mid \mathbf{X}_{ik}) = \mathbf{K}^T\mathbf{A}_i\boldsymbol{\beta} - \mathbf{K}^T\boldsymbol{\theta}\boldsymbol{\phi} + \mathbf{K}^T\boldsymbol{\theta}\mathbf{X}_{ik}.$$

Using (4.2), we see that the regression coefficient for the covariates reduces to a constant for the appropriate functions of the covariates:

$$\begin{aligned}\mathbf{K}^T\boldsymbol{\theta}\mathbf{X}_{ik} &= \frac{\sigma_{xy}}{\sigma_{xx}}\mathbf{K}^T\mathbf{I}_p\mathbf{X}_{ik} + \frac{\eta_{xy}\sigma_{xx} - \eta_{xx}\sigma_{xy}}{\sigma_{xx}(\sigma_{xx} + p\eta_{xx})}\mathbf{K}^T\mathbf{J}_p\mathbf{X}_{ik}\\ &= \frac{\sigma_{xy}}{\sigma_{xx}}\mathbf{K}^T\mathbf{X}_{ik}\\ &= \theta_{\mathrm{W}}\mathbf{X}_{ik}^{\mathrm{W}}, \text{ say.} \qquad (4.3)\end{aligned}$$

This implies that a conventional (i.e. each response adjusted by its associated covariate only) within-subject analysis of covariance will be unbiased for the treatment effects. Such an analysis would be produced, for example, by using fixed subject effects with the original data.

We now consider the between-subject information, which can be represented by $\mathbf{j}_p^T\mathbf{Y}_{ik}$. This has regression model

$$\mathrm{E}(\mathbf{j}_p^T\mathbf{Y}_{ik} \mid \mathbf{X}_{ik}) = \mathbf{j}_p^T\mathbf{A}_i\boldsymbol{\beta} - \mathbf{j}_p^T\boldsymbol{\theta}\boldsymbol{\phi} + \mathbf{j}_p^T\boldsymbol{\theta}\mathbf{X}_{ik},$$

and the regression coefficient of the covariates reduces to

$$\begin{aligned}\mathbf{j}_p^T\boldsymbol{\theta}\mathbf{X}_{ik} &= \frac{\sigma_{xy}}{\sigma_{xx}}\mathbf{j}_p^T\mathbf{I}_p\mathbf{X}_{ik} + \frac{\eta_{xy}\sigma_{xx} - \eta_{xx}\sigma_{xy}}{\sigma_{xx}(\sigma_{xx} + p\eta_{xx})}\mathbf{j}_p^T\mathbf{J}_p\mathbf{X}_{ik} \\ &= \frac{\sigma_{xy} + p\eta_{xy}}{\sigma_{xx} + p\eta_{xx}}\mathbf{j}^T\mathbf{X}_{ik} = \theta_{\mathrm{B}}X_{ik}^{\mathrm{B}}, \text{ say.}\end{aligned} \tag{4.4}$$

This implies that, for designs in which treatments can be estimated using between-subject information only, such as the main plot treatment in a split-plot design, these treatment estimators will be unbiased.

However, when both between-subject and within-subject information are combined in the conventional REML analysis of covariance, the assumption is implicitly made that both the within-subject covariate functions $\mathbf{X}_{ik}^{\mathrm{W}}$ and between-subject function X_{ik}^{B} have the same regression coefficient, that is, $\theta_{\mathrm{W}} = \theta_{\mathrm{B}}$, and it is clear from (4.3) and (4.4) that this can never be true unless the very artificial constraint $\sigma_{xy} = \eta_{xy} = 0$ holds. Thus, in general, the conventional analysis of covariance will be biased for the treatment effects. This is an example of so-called "cross-level bias".

This bias can be avoided in 2 ways. In the first, the analysis is restricted to within-subject information, that is, by using fixed subject effects. This is anyway the appropriate approach when there is little or no relevant between-subject information in the data, such as with a Williams design. Second, if random subject effects are used, then different coefficients must be allowed for the between- and within-subject covariate regressions. This can be done simply by adding to the model the variate that consists of the value of the covariate averaged over each subject. An alternative solution under the random subject effects model is to introduce fixed sequence effects, but as this anyway removes all between-subject information from the treatment estimators, it is a rather pointless exercise; it is then more logical to use fixed subject effects, or use the baselines as outcomes.

We now return to the 2 illustrative examples. First, we consider the balanced, orthogonal Williams design from the study on blood pressure. In the absence of covariates, treatment effects are, in this design, completely orthogonal to periods and subjects. The introduction of covariates will introduce some nonorthogonality, but in a design like this it would not be expected to be great. Consequently, within-subject information will still dominate the treatment estimates and so it is the size of the within-subject covariate coefficient that will be critical when considering the impact of wrongly omitting the separation of within-subject and between-subject covariate regressions. When a common covariate is fitted (we label this coefficient θ_{C}), the estimated coefficient will be a weighted combination of the within-subject and between-subject coefficients (labeled θ_{W} and θ_{B}, respectively in (4.3) and (4.4)) and the greater the relative precision of the former, the smaller the difference will be between the $\hat{\theta}_{\mathrm{C}}$ and $\hat{\theta}_{\mathrm{W}}$, which in turn will reduce the resulting bias of the common covariate analysis. For this example, fitting a single overall covariate results in an estimate (SE) of $\hat{\theta}_{\mathrm{C}} = 0.08$ (0.23), which is close to zero. However, the within-subject and between-subject estimates are $\hat{\theta}_{\mathrm{W}} = -0.23$ (0.24) and $\hat{\theta}_{\mathrm{B}} = 0.78$ (0.31), respectively. In terms of precision, the within-subject estimate remains negligible, but the between-subject coefficient is comparatively large and statistically significant. The consequences for the high–low treatment comparison are as follows. For the common and separate covariate analyses, the estimated comparisons (SEs) are 6.53 (3.18) and 7.04 (3.07), respectively. The impact of the difference in covariate coefficient is, in the present example, not great, being equivalent to an increase in sample size of 7%. This is partly because the analysis largely involves only within-subject information; it would typically be considerably greater in a highly nonorthogonal design. The impact is however sufficient in this example to change the associated P-value from $P = 0.06$ to $P = 0.03$ using for the analyses common and separate covariates, respectively. The impact also depends however on the imbalance of the covariates between the treatments within periods: if they are perfectly balanced, the covariate has no effect on the adjusted means. Here, the imbalance is comparatively small.

For the second illustrative example, it is much harder to predict the impact of inappropriately using the common covariate. Within-subject and between-subject information is combined in a complex way both in estimating the effects of interest and in estimating the covariate coefficients, and the lack of balance means that, in principle, the consequences can be quite different for different treatment effects. In this particular example, the very large subject-level correlation between baseline and response noted earlier (0.997) will have a large impact. Fitting a common covariate produces an estimated coefficient of $\hat{\theta}_{\mathrm{C}} = 1.01$ (0.03), while fitting the covariate separately at the within-subject and between-subject levels produces estimated coefficients of $\hat{\theta}_{\mathrm{W}} = 0.42$ (0.10) and $\hat{\theta}_{\mathrm{B}} = 1.04$(0.03), respectively. The latter coefficient is very close to one and is far more precise than the within-subject coefficient, and that the combined estimate $\hat{\theta}_{\mathrm{C}}$ is, as seen, also close to one, implying that adjustment by the single covariate and change from baseline will, in this very particular setting, be numerically very similar, implying in turn that the bias should not be great using the single covariate in spite of the large difference between $\hat{\theta}_{\mathrm{W}}$ and $\hat{\theta}_{\mathrm{B}}$. This is indeed what is seen. For comparison between the positive and negative controls, the use of the single covariate produces an estimate SE of 0.225 (0.055) which is almost identical to that seen in Section 4.1 from the analysis of the change from baseline D: 0.224 (0.055). Adjustment using both within-subject and between-subject covariates produces by comparison an estimate SE of 0.173 (0.046). There is an increase in precision associated with the smaller estimated within-subject residual, and the change in estimated coefficient is of the order of 1 SE. For the comparison between high dose and low dose, the corresponding 3 estimates SEs are 0.011 (0.100), 0.014 (0.099), and 0.040 (0.084). The relative decrease in the size of the standard error is the same for the 2 comparisons. But the change in estimated value is less extreme in the second case than in the first because this comparison makes more use of the between-subject information. The gain in precision through using the 2 baseline covariates is equivalent to a change in sample size of 40%.

4.3 *Baselines as response variables*

An alternative approach to inclusion of the baseline variables in the analysis is to treat them as additional response variables without accompanying fixed treatment effects. Such an approach is well established in the analysis of parallel group longitudinal studies (see e.g. Carpenter and Kenward, 2008, Chapter 3). One important result is that in particular balanced orthogonal settings under certain models, the treatment estimates obtained through the use of baselines as covariates or as responses are identical, with very similar standard errors and a difference of one in degrees of freedom. A proof of this is given in Carpenter and Kenward (2008, Appendix 4.4). In other settings, such as with unbalanced and/or nonorthogonal designs, we typically expect very similar estimates and inferences that converge asymptotically. There will be systematic differences between the 2 approaches, typically still small however, when numbers of observations differ among subjects. We return to this issue in Section 4.4. Here, we assume that there are no missing values and that numbers of periods are the same for all subjects.

The model, we use is precisely the joint one for the time-ordered data $\mathbf{Z}_{ik}$ described in Section 2, and can be fitted using the generic SAS MIXED code given in the Appendix. Note that we do not in general constrain the period effects associated with baseline and response to be the same, hence the presence of the type-by-period interaction in the MIXED model statement.

Applying this analysis to the first example from the blood pressure trial, we obtain an estimated high- versus low-dose comparison (SE, degrees of freedom [DF]) of $-0.949(1.793, 20)$, compared with that from the analysis of covariance in Section 4.3 with separate within-subject and between-subject covariates of $-0.949(1.763, 19)$. In this very special balanced/orthogonal setting, we see, as predicted, that the estimates are identical, the SEs are very similar, and the DF differ by one. With such a design, and complete data, both analyses are, for practical purposes, identical.

The second example is neither balanced nor orthogonal. We begin with the first contrast, positive versus negative control. Applying the above approach to the joint analysis of the baselines and outcomes we get a contrast estimate (SE, DF) of 0.170 (0.045, 67.6), compared with the analysis of covariance from Section 4.3: 0.173 (0.046, 66.1). The estimates and associated SEs are similar but the difference in DF suggests that more is different between these 2 analyses than in the balanced/orthogonal case. This is indeed the case, and the difference is partly due to the missing data. We return to this in Section 4.4. A similar picture is seen with the second treatment contrast, the high- versus low-dose comparison. The estimates (SE, DF) from the joint and covariance based analyses are 0.0343 (0.0825, 82) and 0.0270 (0.0830, 83.4), respectively. Here, the estimates show a greater difference. The patterns in these 2 sets of comparisons reflect the range of influences on differences between these 2 types of analysis: lack of balance, nonorthogonality, incompleteness, and estimated variance parameters. In spite of these however, the differences remain small, and are negligible from a practical perspective.

We note finally that, strictly, the sample space is different in the 2 types of analysis, respectively marginal and conditional, for the joint model and covariate-based analysis. This implies that we should be careful in comparing, for example, precision from the 2 approaches. We see however, both from theory and empirically, that in such settings the practical differences between the analyses in terms of estimates, precision, and inferences are usually negligible, and so we do not regard this distinction as an important issue in the current setting.

4.4 *Incomplete data*

As pointed out in Sections 1 to 4.2, without specific constraints on the covariance structure of the $2p$ repeated measurements, all p baselines $\mathbf{X}_{ik}$ will appear as covariates in the analysis of covariance of $\mathbf{Y}_{ik}$, and the coefficients of these may be different among the Ys within a subject. The reduction of these to just 2 common coefficients for the within-subject baselines ($\mathbf{X}_{ik}^{\mathrm{W}}$) and for the average baseline (X_{ik}^{B}) depends on the particular form of covariance structure specified in (2.5), with the relevant coefficients presented in (4.3) and (4.4), respectively. For the derivation of these, and the consequences on the analysis, it was assumed in Section 4.2 that p was a constant, that is, that all subjects had a complete set of measurements. When data are missing, which is a common occurrence in practice, this will no longer be the case and we examine here, the implications of this for the results presented so far. Because of the symmetry of our covariance structure, the implication of missing data will be the same (in terms of the covariate structure of the analysis of covariance) from whichever periods the data are missing. So it is only the total number of periods observed for a particular subject, p_{ik} say, that is relevant, not the pattern of missingness, and we can apply the arguments leading to (4.3) and (4.4) subject-by-subject, replacing p by p_{ik} in each case. It follows immediately that the reduction to the within-subject and between-baseline covariates, ($\mathbf{X}_{ik}^{\mathrm{W}}$ and X_{ik}^{B}), still holds. Next, we note that the coefficient for $\mathbf{X}_{ik}^{\mathrm{W}}$ does not depend on p, so this remains the same irrespective of missing data, and therefore holds across all subjects. In contrast, we see that the coefficient of X_{ik}^{B} does depend on p. The coefficient of this covariate therefore depends on the number of missing data and, strictly, in the analysis of covariance should be allowed to have different values for each of the possible values of p. Holding this to a constant value, as we have done above in the analysis of the second example introduces bias into the estimates which does not disappear with increasing sample size if the proportion and pattern of missing data is maintained. We conjecture, however, than in practice such bias will be very small indeed, and the simpler analysis presented earlier will provide a perfectly acceptable approximation. There are 2 main reasons for this. First, in the simpler analysis, the bias only comes from the contribution of the between-subject information, which is typically the much less important component. Second, unless an unusually high proportion of subjects have missing data, the estimate of the common covariate coefficient will anyway be dominated by subjects with complete data, and so be close to the required coefficient for these subjects. Finally, in the approach of Section 4.4 in which the baselines are treated as responses, this modification of the between-subject covariate coefficient is (implicitly) done correctly and this source of bias does not arise.

Only the second illustrative example has missing data, but there are 2 main reasons why we expect the bias discussed here to be negligible in this case. First, only 4 subjects have missing data. Second, the between-subject covariance parameters (η_{xx}, η_{yy}, and η_{xy}) dominate the within-subject parameters (σ_{xx}, σ_{yy}, and σ_{xy}) and, if the coefficient (4.4) is re-written as

$$\theta_{\mathrm{B}} = \frac{\eta_{xy} + \sigma_{xy}/p}{\eta_{xx} + \sigma_{xx}/p},$$

it is clear that in such circumstances the coefficient is almost independent of p. We see this reflected in the analysis. Looking at the first treatment comparison, positive versus negative control, the previous analysis of covariance gave an estimate (SE, DF) of 0.173 (0.046, 66.1). Allowing θ_{B} to differ for the subject with incomplete data gives, by comparison, 0.170 (0.045, 67.6), which is very similar.

In conclusion, strictly, when there are missing data, the between-subject covariate coefficient should be allowed to differ according to the size of p_{ik}. We suggest, however, that, in practice, the simpler analysis with common coefficient will provide a very good approximation to such approach. If there are concerns about such bias, the coefficient can be allowed to differ in the analysis, or the marginal analysis of Section 4.3 can be used.

Finally, we note that when baselines are missing but their associated response variables are observed, the joint analysis of Section 4.3 does have an advantage over the analysis of covariance. The analysis of covariance will discard the associated responses, while the joint approach will lead to their inclusion.

5. Discussion

A crossover study generates repeated measurements data which, from the second period onward, have much in common with data from an observational study. In particular, baseline measurements made at the start of treatment periods beyond the first are postrandomization observations. Using the framework of causal inference (e.g. Greenland *and others*, 1999), we would say that these later baselines lie on the causal pathway, and as a consequence great care needs to be made when inferences are made conditionally upon them. This is a distinct issue to that of carryover, which we have been assuming throughout to be negligible. The problem can also be seen as one of estimating separate regression coefficients at 2 structural levels in a hierarchical model as has been considered in the past by a number of authors. For example, Goldstein (1987, Section 3.7) discusses it in the context of survey data, Skrondal and Rabe-Hesketh (2004, Chapter 3) introduce it in the more general problem of endogeneity with reference to econometrics, and Galwey (2006, Section 7.4) presents an example in an experimental setting.

To explore the use of such covariates in practice, we have postulated a 6-parameter covariance structure for repeated measurements from a crossover trial that reflects the exchangeability one would expect to find but, at the same time, does not impose unnecessary constraints on this structure. Several empirical examples of the fitted covariance structure are presented.

In a crossover setting a conventional analysis of covariance mimics the one that would apply to data from a parallel-group study: each baseline is used as a covariate for the following outcome and for no other. We have shown that such an analysis will lead to biased treatment comparisons in all but the most contrived and unrealistic settings if random subject effects are used. In many crossover trials, we expect the between-subject correlation to be very large. The remaining within-period (or local) correlation will typically be much smaller and in many cases quite close to zero. This means that the regression coefficient for baseline from the single baseline covariate analysis (θ_C in our notation) will be overestimated and subsequently the adjustment of treatment effects for baseline imbalance will be exaggerated, leading to bias. This will be especially true in simple orthogonal designs such as Williams squares where the treatment differences are estimated at the within-subject level. We show that this bias is an example of so-called cross-level bias (e.g. Sheppard, 2003), and we quantify it in terms of the design and the 6 parameters of our covariance structure. We demonstrate how the bias can be removed in 3 different ways. First, fixed subject effects can be used. This removes all between-subject information from the analysis, which will have a nonnegligible impact on precision in all but the most inefficient of designs. Second, if random effects are to be retained, then separate regression coefficients are required for within-period and between-period (subject-average) covariates. Strictly, a small refinement to this is required when subjects have different numbers of measurements, due to dropout, for example, but ignoring this is likely to introduce only a nugatory bias. A modification of these 2 approaches uses random subject effects together with fixed sequence effects. We do not recommend this approach because it is essentially self-contradictory: random effects are included to allow the incorporation of between-subject information, while fixed sequence effects remove this information. In our third approach, the baselines are also treated as outcome variables but without associated treatment effects. This route has the advantage of automatically producing the correct analysis where there are missing data, that is, avoiding the approximation implicit in the first method, and allows the incorporation of outcomes with missing associated baselines.

The saving in terms of precision through the appropriate use of baseline measurements can, in certain settings, be quite large, depending on the particular design and outcome variable. In one of our examples, this gain was equivalent to a 40% increase in sample size.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org

ACKNOWLEDGMENTS

We are grateful to a referee for helpful comments on an earlier version of this paper. *Conflict of Interest:* None declared.

APPENDIX

The following SAS PROC MIXED code will fit the model in (2.1) using the covariance structure given by (2.5) expressed in the form (2.6). We define the following variates: `Subjid`, the subject identifier; `Pernum`, the period label, with baseline and outcome sharing the same period; `Group`, the treatment identifier associated with the outcome variable with a separate level (zero, say) for baseline; `Type`, an indicator variable taking the value 1 for outcome (Y) and 0 for baseline (X). The data, `Z`, consist of the baselines and outcomes in time order.

```
proc mixed data = RM;
class Subjid Pernum Group Type;
model Z = Pernum*Type Group*Type /solution ddfm = kr;
random Type /subject = Subjid Type = UN;
repeated Type /Subject = Subjid*Pernum Type = UN;
run;
```

References

Carpenter, J. R. and Kenward, M. G. (2008). *Missing Data in Randomised Controlled Trials—A Practical Guide*. http://www.pcpoh.bham.ac.uk/publichealth/nccrm/PDFs_and_documents/Publications/Final_Report_RM04_JH17_mk.pdf.

Freeman, P. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in Medicine* **8**, 1421–1432.

Galwey, (2006). *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance.* Chichester: Wiley.

Goldstein, H. (1987). *Multilevel Models.* London: Charles Griffin & Co.

Greenland, S., Pearl, J. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48.

Jones, B. and Kenward, M. G. (2003). *Design and Analysis of Cross-over Trials,* 2nd edition. London: Chapman & Hall/CRC.

Kenward, M. G. and Jones, B. (1987). The analysis of data from 2x2 cross-over trials with baseline measurements. *Statistics in Medicine* **6**, 911–926.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects estimators from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Kenward, M. G. and Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis* **53**, 2583–2595.

McCann, D., Barrett, A., Cooper, A., Crumpler, D., Dalen, L., Grimshaw, K., Kitchin, E., Lok, K., Porteous, L., Prince, E. *and others* (2007). Food additives and hyperactive behaviour in 3-year-old and 8/9-year-old children in the community: a randomised, double-blinded, placebo-controlled trial. *The Lancet* **370**, 1560–1567.

Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: curent practice and problems. *Statistics in Medicine* **21**, 2917–2930.

Senn, S. J. (1988). Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine* **7**, 1099–1101.

Senn, S. J. (2002) *Cross-over Trials in Clinical Research,* 2nd edition. Chichester: Wiley.

Sheppard, L. (2003) Insights on bias and information in group-level studies. *Biostatistics* **4**, 265–278.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling.* London: Chapman & Hall/CRC.

Stylianou, A., Roger, J. and Stephens, K. (2008). A statistical assessment of QT data following placebo and moxifloxacin dosing in thorough QR studies. *Journal of Biopharmaceutical Statistics* **18**, 502–516.

[*Received March 27, 2009; revised July 15, 2009; accepted for publication October 7, 2009*]