STATISTICS IN MEDICINE, VOL. 11, 1685–1704 (1992)

REPEATED MEASURES IN CLINICAL TRIALS: ANALYSIS USING MEAN SUMMARY STATISTICS AND ITS IMPLICATIONS FOR DESIGN

LARS FRISON* AND STUART J. POCOCK

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, U.K.

SUMMARY

This paper explores the use of simple summary statistics for analysing repeated measurements in randomized clinical trials with two treatments. Quite often the data for each patient may be effectively summarized by a pre-treatment mean and a post-treatment mean. Analysis of covariance is the method of choice and its superiority over analysis of post-treatment means or analysis of mean changes is quantified, as regards both reduced variance and avoidance of bias, using a simple model for the covariance structure between time points. Quantitative consideration is also given to practical issues in the design of repeated measures studies: the merits of having more than one pre-treatment measurement are demonstrated, and methods for determining sample sizes in repeated measures designs are provided. Several examples from clinical trials are presented, and broad practical recommendations are made. The examples support the value of the compound symmetry assumption as a realistic simplification in quantitative planning of repeated measures trials. The analysis using summary statistics makes no such assumption. However, allowance in design for alternative non-equal correlation structures can and should be made when necessary.

1. INTRODUCTION

In clinical trials and other experimental studies one commonly encounters repeated observations of a quantitative outcome measure on every subject at several pre-defined times since randomization. Such post-randomization repeated measurements over time are often accompanied by one or more pre-randomization (baseline) measurements on each subject.

There exists a battery of relatively complex methods for analysing such data (for example repeated measures ANOVA, MANOVA and multi-level models) and these are described in textbooks such as those by Crowder and Hand,[1] Goldstein,[2] and Milliken.[3] In most medical journals such complex methods are rarely used perhaps because they are difficult to communicate to non-statisticians.[4] However, some specialist journals (especially those in psychiatry) seem to encourage repeated measures (split-plot in time) ANOVA, even though it has been called a 'dangerously wrong' method.[5] A survey of published clinical trials in general medical journals[6] found that reported analyses of repeated measurement data were mostly either purely descriptive or misguidedly entailed repeated use of significance tests at every time point without correction for multiple testing.

Matthews *et al.*,[7] in a valuable contribution to bridging the gap between inadequate statistical practice and complex statistical theory, argue for the use of summary measures to analyse

* Currently at Astra Hassle, Molndal, Sweden. Correspondence to Stuart Pocock.

0277–6715/92/131685–20$15.00

Received October 1991
Revised May 1992

repeated measures. The intention is to decide in advance on an appropriate summary of each individual's response to treatment, and then to use simple two-group comparison techniques, such as the two-sample t-test, to assess treatment differences in the summary measures.[8] An alternative approach is to consider repeated measures as a multiple end-points problem, for which various methods of accounting for multiple testing have been developed,[9,10] but we feel this will usually be less informative than a summary statistic approach.

The choice of summary measures is wide, for example the post-treatment mean, mean change relative to baseline, end value, end value minus baseline, slope, maximum value, area under the curve, time to reach a peak or a nominated value. However, in many clinical trials the prime objective is to assess the average response to treatment over time, often (but not necessarily) in anticipation that treatment response is liable to occur quickly and to remain reasonably steady over time. This provides three possible methods of analysis:

1. *Post-treatment means* (POST): a simple analysis using the mean for each patient's post-treatment measurements as the summary measure.
2. *Mean changes* (CHANGE): a simple analysis of each patient's difference between mean of post-treatment measurements and mean of baseline measurements, the latter often consisting of just a single baseline value per patient.
3. *Analysis of covariance* (ANCOVA): between-patient variations in baseline measurements are taken into account, by using the mean baseline measurement for each patient as a covariate in a linear model for treatment comparison of post-treatment means.

For brevity, these methods will henceforth be referred to as POST, CHANGE and ANCOVA respectively.

In Section 2 we define a simple model for randomized trials with repeated measures, and illustrate its value with an example. Section 3 explores the statistical properties of these three commonly used approaches, in particular documenting the superiority of ANCOVA. In Section 4 a more precise quantitative comparison of variances for the three estimation methods is made for the common case of a single pre-treatment measurement. While the three methods can be formulated as significance tests (two-sample t-tests and a covariate-adjusted test of difference in mean respectively) we prefer to emphasize their use in estimating treatment effects. Also, our results apply more directly to trials with reasonable sample sizes.

We have found little published information on statistical design considerations in repeated measurements studies. Hence, in Sections 5 and 6 we focus on how to choose the numbers of baseline and post-treatment measurements, and use of power calculations for determining the required number of patients in repeated measures designs. The extent of bias in estimation if ANCOVA is not used is described in Section 7. For simplicity of illustration, Sections 4 to 6 assume compound symmetry (identical within-group variances over time and identical correlations between all pairs of measurements). Section 8 presents analyses of the example introduced in Section 2. Also, a variety of examples of clinical trials are used to assess correlation structures and the practical validity of the compound symmetry assumption. Section 9 evaluates the effect of unequal correlations on design considerations. Finally in Section 10 we discuss the value and limitations of these relatively simple approaches in design and analysis of clinical trials with repeated measurements over time.

2. A SIMPLE MODEL AND AN EXAMPLE

Suppose a randomized clinical trial has two treatment groups ($i = A$ or B) with n_i patients per group, and suppose all patients have p pre-treatment visits, $k = -(p-1), \ldots, 0$, and r post-

treatment visits, $k = 1, \ldots, r$. A quantitative measurement x is observed at every visit for every patient, and we adopt the simple model:

$$x_{ijk} = \mu_{ik} + e_{ijk} \qquad \text{for } i = A \text{ or } B, j = 1, \ldots, n_i \text{ and } k = -(p-1), \ldots, 0, 1, \ldots, r,$$

Here μ_{ik} is the underlying true mean response for treatment i at time k. As a result of randomization we can assume $\mu_{Ak} = \mu_{Bk}$ for the pre-treatment visits $k \leqslant 0$. e_{ijk} is the individual jth patient 'error' or variation around the underlying mean μ_{ik}, and these errors will not be independent within patients.

Hence, let $\Sigma = \{\sigma_{kl}\}$ be the covariance matrix for all pairs of measurement times k, l. For simplicity we assume this is the same for both treatments. It is helpful to define three submatrices:

$$\Sigma_{\text{post}} = \{\sigma_{kl}\} \qquad \text{for} \quad k, l = 1, \ldots, r;$$

$$\Sigma_{\text{pre}} = \{\sigma_{kl}\} \qquad \text{for} \quad k, l = -(p-1), \ldots, 0;$$

and

$$\Sigma_{\text{mix}} = \{\sigma_{kl}\} \qquad \text{for} \quad k = -(p-1), \ldots, 0 \qquad \text{and} \quad l = 1, \ldots, r.$$

Thus we can display

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{\text{pre}} & \Sigma'_{\text{mix}} \\ \hline \Sigma_{\text{mix}} & \Sigma_{\text{post}} \end{array}\right].$$

It is also convenient to define $\sigma_{kl} = \rho_{kl}\sigma_k\sigma_l$, where ρ_{kl} is the within-treatment group pairwise correlation between a patient's measurements at visits k and l, and σ_k, σ_l are the standard deviations at visits k, l within each treatment group. We expect the correlations ρ_{kl} to be substantial (typically greater than 0·5 in most trials) since they reflect the consistency of patient effects over time, which are otherwise not explicitly included in this simple model.

We now illustrate the value of this model with a practical example. A randomized trial of 152 patients with coronary heart disease compared an active drug with a placebo during a 12 month follow-up period. The liver enzyme CPK in serum was measured to study a possible adverse drug effect on the liver. Each patient had *three pre-treatment measurements*, taken 2 months before, 1 month before and at randomization, and *eight post-treatment measurements*, taken every 1·5 months after randomization.

Figure 1 shows the results as commonly displayed in a medical journal, with means by treatment group for every time point. While there is a consistent pattern of higher post-treatment means on the active drug, the standard errors are substantial. The common but misguided practice of separate significance testing for each post-treatment time point reveals a varied collection of t-statistics, whether we use means, mean changes or ANCOVA. The t-values range from 0·35 (ANCOVA for visit 12 with visit 0 as covariate) to 3·34 (ANCOVA for visit 4·5 with mean of visits -2, -1 and 0 as covariate) with around half the time-point-specific significance tests having $P < 0{\cdot}05$ whichever method of analysis is used.

However, this plethora of significance tests is based on the false premise that each time point is of separate interest in its own right. In reality, the primary hypothesis is more global (across all post-treatment measurements, is there a tendency for an elevation in CPK on the active drug?) and in Section 8 we return to this example with appropriate analyses.

In exploring the correlation structure in these data, each pairwise correlation ρ_{kl} has been estimated by $\hat{\rho}_{kl}$, the observed correlations obtained from a weighted average of the two

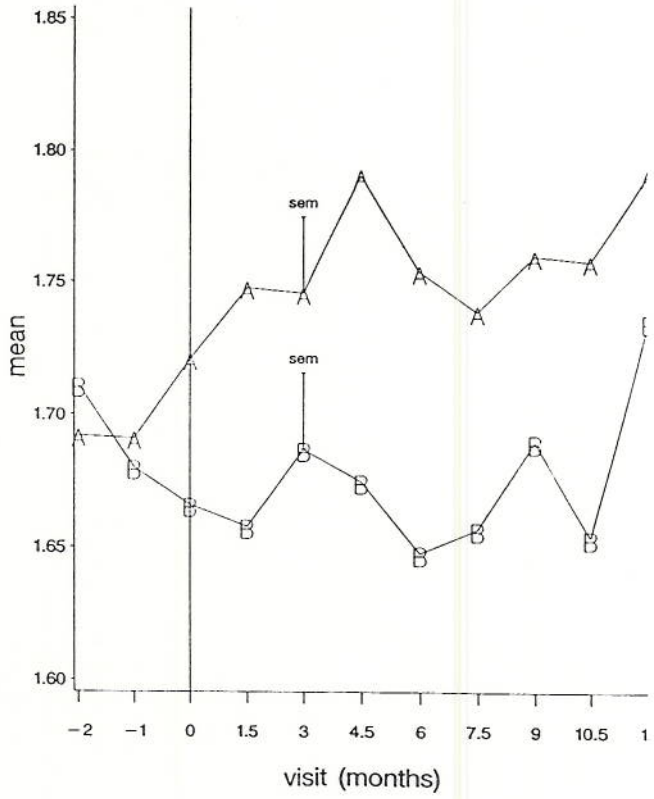

Figure 1. Mean level of CPK over time for drug A ($n = 76$) and drug B ($n = 76$): standard error of mean (SEM) shown only for 3 month visit, others are of similar magnitude

treatment groups' covariance matrices, weights being proportional to sample size. Figure 2 plots $\hat{\rho}_{kl}$ by the time between visits k and l; pre–pre, pre–post and post–post pairs are denoted by different symbols. There is a general consistency in the correlations, all being in the range 0·5 to 0·8. Also, the three types of pairs show similar magnitude. There is a slight decline in correlation amongst more distant pairs of time points, the estimated slope being $-0{\cdot}009$ per month apart. This indicates only slight departure from the assumption that ρ_{kl} is constant for any $k \neq l$. Also, the variance σ_k^2 varied little between visits. This assumption of compound symmetry is useful in several sections to follow. In Section 8 we study the correlation structure of several repeated measures trials to explore its feasibility in practice and to assess the likely magnitude of correlations in general.

3. VARIANCE FORMULAE FOR THREE ANALYSIS APPROACHES

Let $\bar{\Sigma}_{\text{post}}$, $\bar{\Sigma}_{\text{pre}}$ and $\bar{\Sigma}_{\text{mix}}$ be the respective means of the r^2, p^2 and $r \times p$ components of the three submatrices Σ_{post}, Σ_{pre} and Σ_{mix} defined in Section 2. Using this notation we can define the following variance formulae for the three analysis approaches defined in Section 1.

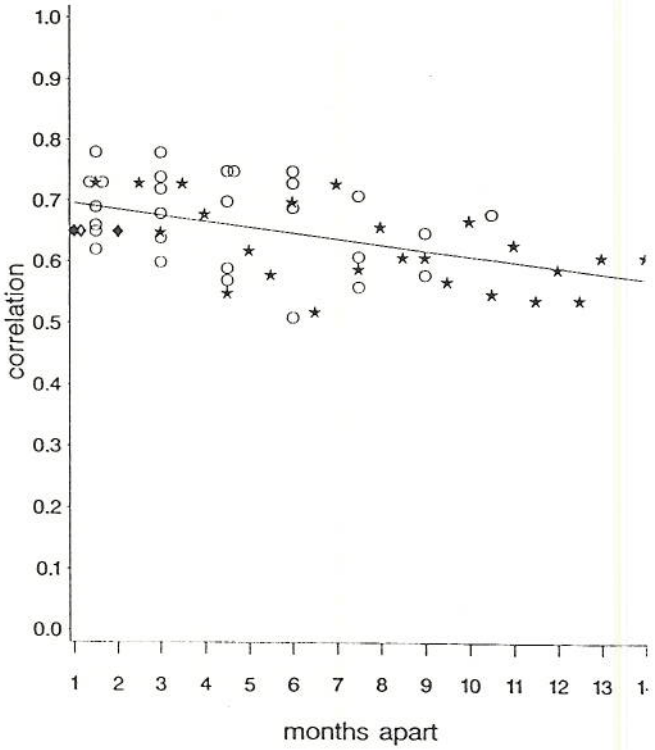

Figure 2. CPK, $n = 152$, correlation coefficients versus time between visits: ◆ = pre–pre, ★ = pre–post, ○ = post–post

Post-treatment means (POST)

For each individual the summary statistic is

$$\bar{x}_{ij.}^{\text{post}} = \frac{1}{r}\sum_{k=1}^{r} x_{ijk}.$$

The overall post-treatment mean difference is

$$\frac{1}{n_A}\sum_{j=1}^{n_A} \bar{x}_{Aj.}^{\text{post}} - \frac{1}{n_B}\sum_{j=1}^{n_B} \bar{x}_{Bj.}^{\text{post}} = \bar{x}_{A..}^{\text{post}} - \bar{x}_{B..}^{\text{post}},$$

which has expected value

$$\frac{1}{r}\sum_{k=1}^{r} (\mu_{Ak} - \mu_{Bk}) = \bar{\mu}_{A.}^{\text{post}} - \bar{\mu}_{B.}^{\text{post}}.$$

It is easily shown that

$$\text{var}(\bar{x}_{A..}^{\text{post}} - \bar{x}_{B..}^{\text{post}}) = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\bar{\Sigma}_{\text{post}}.$$

Mean changes (CHANGE)

For each individual the summary statistic is the mean change,

$$\frac{1}{r}\sum_{k=1}^{r} x_{ijk} - \frac{1}{p}\sum_{k=-(p-1)}^{0} x_{ijk} = \bar{x}_{ij.}^{\text{post}} - \bar{x}_{ij.}^{\text{pre}}.$$

Then the overall treatment difference in these mean changes is

$$\frac{1}{n_A}\sum_{j=1}^{n_A}(\bar{x}_{Aj.}^{\text{post}} - \bar{x}_{Aj.}^{\text{pre}}) - \frac{1}{n_B}\sum_{j=1}^{n_B}(\bar{x}_{Bj.}^{\text{post}} - \bar{x}_{Bj.}^{\text{pre}}) = (\bar{x}_{A..}^{\text{post}} - \bar{x}_{A..}^{\text{pre}}) - (\bar{x}_{B..}^{\text{post}} - \bar{x}_{B..}^{\text{pre}}),$$

which has expected value again equal to $\bar{\mu}_A^{\text{post}} - \bar{\mu}_B^{\text{post}}$ since the pre-treatment expected values are the same for both treatments. It is easily shown that

$$\text{var}[(\bar{x}_{A..}^{\text{post}} - \bar{x}_{A..}^{\text{pre}}) - (\bar{x}_{B..}^{\text{post}} - \bar{x}_{B..}^{\text{pre}})] = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)(\bar{\Sigma}_{\text{post}} + \bar{\Sigma}_{\text{pre}} - 2\bar{\Sigma}_{\text{mix}}).$$

Analysis of covariance (ANCOVA)

The model for ANCOVA based on the individual's post-treatment mean $\bar{x}_{ij.}^{\text{post}}$, with the pre-treatment mean $\bar{x}_{ij.}^{\text{pre}}$ as a covariate, is a follows:

$$\bar{x}_{ij.}^{\text{post}} = \bar{\mu}_{i.}^{\text{post}} + \beta(\bar{x}_{ij.}^{\text{pre}} - \bar{\mu}_{..}^{\text{pre}}) + \varepsilon_{ij},$$

where ε_{ij} are independent random errors with assumed constant variance. With estimate $\hat{\beta}$ obtained by least squares, we may define $\bar{x}_{ij.}^{\text{cov}} = \bar{x}_{ij.}^{\text{post}} - \hat{\beta}(\bar{x}_{ij.}^{\text{pre}} - \bar{x}_{..}^{\text{pre}})$. Then the estimated mean treatment difference is

$$\frac{1}{n_A}\sum_{j=1}^{n_A}\bar{x}_{ij.}^{\text{cov}} - \frac{1}{n_B}\sum_{j=1}^{n_B}\bar{x}_{ij.}^{\text{cov}} = \bar{x}_A^{\text{cov}} - \bar{x}_B^{\text{cov}},$$

which again has expected value $\bar{\mu}_{A.}^{\text{post}} - \bar{\mu}_{B.}^{\text{post}}$. From the variance formula for ANCOVA,[11]

$$\text{var}(\bar{x}_A^{\text{cov}} - \bar{x}_B^{\text{cov}}) = \frac{n_A + n_B - 2}{n_A + n_B - 3}[1 - \text{corr}^2(\bar{x}_{ij.}^{\text{pre}}, \bar{x}_{ij.}^{\text{post}})] \times \text{var}(\bar{x}_{ij.}^{\text{post}})\left[\frac{1}{n_A} + \frac{1}{n_B} + \frac{(\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}})^2}{(n_A + n_B - 2)\,\text{var}(\bar{x}_{ij.}^{\text{pre}})}\right].$$

The first term corresponds to the loss of one degree of freedom, due to estimation of the slope. The additional correction factor in the last term allows for the fact that the sampling error of the estimated slope leads to a correlation between $\bar{x}_A^{\text{cov}}$ and $\bar{x}_B^{\text{cov}}$.

Using the above notation for components of Σ,

$$\text{var}(\bar{x}_A^{\text{cov}} - \bar{x}_B^{\text{cov}}) = \frac{n_A + n_B - 2}{n_A + n_B - 3}\left(\bar{\Sigma}_{\text{post}} - \frac{\bar{\Sigma}_{\text{mix}}^2}{\bar{\Sigma}_{\text{pre}}}\right)\left[\frac{1}{n_A} + \frac{1}{n_B} + \frac{(\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}})^2}{(n_A + n_B - 2)\bar{\Sigma}_{\text{pre}}}\right].$$

As the sample size increases the first term approaches unity, and in randomized trials the correction factor in the last term becomes negligible. Hence for any reasonable size of trial we can use the simpler approximation

$$\text{var}(\bar{x}_A^{\text{cov}} - \bar{x}_B^{\text{cov}}) \approx \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\left(\bar{\Sigma}_{\text{post}} - \frac{\bar{\Sigma}_{\text{mix}}^2}{\bar{\Sigma}_{\text{pre}}}\right).$$

In summary, for a randomized clinical trial all three estimates of the mean treatment difference have the same expected value, $\bar{\mu}_A^{post} - \bar{\mu}_B^{post}$. Given the common sample size adjustment $(1/n_A) + (1/n_B)$, the comparison of variance magnitudes may be expressed as:

POST: variance proportional to $\bar{\Sigma}_{post}$
CHANGE: variance proportional to $\bar{\Sigma}_{post} + \bar{\Sigma}_{pre} - 2\bar{\Sigma}_{mix}$
ANCOVA: variance approximately proportional to $\bar{\Sigma}_{post} - (\bar{\Sigma}_{mix}^2/\bar{\Sigma}_{pre})$.

It can be readily seen that ANCOVA always has a smaller variance than POST. ANCOVA also produces a smaller variance than CHANGE, since the difference in variances is proportional to

$$\bar{\Sigma}_{pre} - 2\bar{\Sigma}_{mix} + \frac{\bar{\Sigma}_{mix}^2}{\bar{\Sigma}_{pre}} = \frac{1}{\bar{\Sigma}_{pre}}(\bar{\Sigma}_{pre} - 2\bar{\Sigma}_{mix})^2$$

which cannot be negative. Hence, for this particular summary statistic approach to analysis of repeated measures we confirm the well known result that ANCOVA is superior to both ignoring pre-treatment readings and simply subtracting pre-treatment readings for each individual.[11] In the next two sections we quantify the numerical magnitude of this superiority under specific plausible conditions.

4. COMPARISON OF METHODS WITH A SINGLE PRE-TREATMENT VISIT

Often there is just a single pre-treatment measurement and several (r) measurements after randomization for each patient, and we now focus on this simple case. The general variance formulae in Section 2 enable a comparison of methods for any variance/covariance matrix Σ, but to quantify the differences it is first convenient to display findings under the assumption of compound symmetry.[12] That is, we will assume equal variances for all time points and both treatments and also equal correlations between all pairs of time points. Thus, $\text{var}(x_{ijk}) = \sigma^2$ and $\text{corr}(x_{ijk}, x_{ijl}) = \rho$ for $k \neq l$. While these might appear unrealistic assumptions, our exploration of number of data sets suggest that only modest departures from such compound symmetry are quite common (see Section 8).

Then the three variances can be rewritten as follows:

$$\text{POST:} \quad \text{variance} = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\frac{\sigma^2}{r}[1 + (r-1)\rho]$$

$$\text{CHANGE:} \quad \text{variance} = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\frac{\sigma^2}{r}[1 + (r-1)\rho + r(1-2\rho)]$$

$$\text{ANCOVA:} \quad \text{variance} \approx \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\frac{\sigma^2}{r}[1 + (r-1)\rho - r\rho^2].$$

Figure 3 compares each of the first two approaches with ANCOVA by plotting ratios of variances for various values of ρ and r.

First comparing POST with ANCOVA, the former becomes more inferior the larger the correlation ρ. Also, for any value of ρ this inferiority is somewhat more marked if the number of post-treatment visits r is substantial. If $\rho = 0$, then pre-treatment measurements are of no value, that is $\beta = 0$ in ANCOVA in which case the two approaches are almost equivalent. We may plausibly expect ρ in the range 0·5 to 0·7, in which case the variance for ANCOVA will be around 40 to 60 per cent less than for POST. This reflects the serious loss of statistical efficiency incurred by failing to take account of pre-treatment measurements.

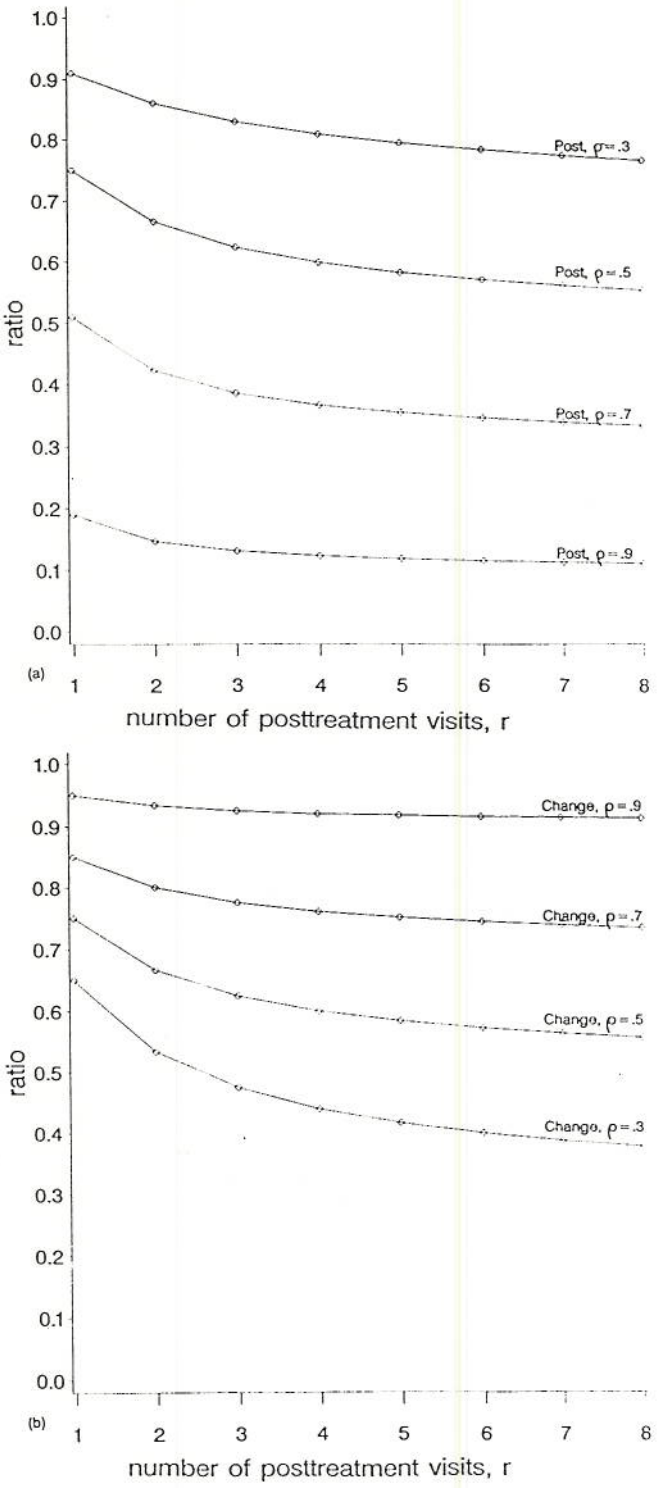

Figure 3. Dependence of (a) var (ANCOVA)/var (POST) and (b) var (ANCOVA)/var (CHANGE) on r and ρ, assuming equi-correlation ρ and one pre-treatment measure

A popular alternative to CHANGE is the percentage change in response. This is readily incorporated into our framework by use of logarithmic transformations, with geometric means used for both individual patient summaries of percentage change and estimates of overall treatment effect. Of course, logarithmic transformations may often be appropriate for POST and ANCOVA as well.

CHANGE becomes less inferior to ANCOVA as the correlation ρ increases. Again, for any value of ρ the inferiority of CHANGE becomes somewhat more accentuated as the number of post-treatment measurements increases. For the plausible values of ρ in the range 0·5 to 0·7, the variance for ANCOVA will be around 20 to 40 per cent less than for CHANGE.

Note that for $\rho = 0{\cdot}5$ POST and CHANGE have identical variances. Our examples (see Section 7 below) suggest ρ will commonly be somewhat higher, so that CHANGE will be better than POST. However, with just a single pre-treatment measurement it seems likely that both these analyses will be substantially inferior to ANCOVA in most practical circumstances.

5. CONSEQUENCES OF HAVING MORE PRE-TREATMENT VISITS

It is often possible to have more than one pre-treatment visit in a repeated measures design (all pre-treatment visits occurring *before* randomization), and here we consider the improved efficiency for both ANCOVA and CHANGE. Of course the time lapses between pre-treatment measurements may affect the correlation structure, but for simplicity we continue to explore statistical properites under the assumption of compound symmetry.

With r post-treatment measurements and p pre-treatment measurements we have

$$\bar{\Sigma}_{\text{post}} = \sigma^2\left[\frac{1+(r-1)\rho}{r}\right], \quad \bar{\Sigma}_{\text{pre}} = \sigma^2\left[\frac{1+(p-1)\rho}{p}\right], \quad \bar{\Sigma}_{\text{mix}} = \sigma^2\rho.$$

For CHANGE:

$$\text{variance} = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\left[\frac{1+(r-1)\rho}{r} - \frac{(p+1)\rho-1}{p}\right]$$

For ANCOVA:

$$\text{variance} \approx \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\left[\frac{1+(r-1)\rho}{r} - \frac{p\rho^2}{1+(p-1)\rho}\right].$$

First, consider the advantage of *extra pre-treatment visits while keeping the number of post-treatment visits fixed*. Then, CHANGE becomes superior to POST provided $\rho > 1/(p+1)$. This means that provision of two or more pre-treatment measurements will make CHANGE the better option unless correlations are small, which appears unlikely in practice.

More important is the extent to which extra pre-treatment measurements make CHANGE closer in statistical efficiency to ANCOVA. From the above formuale it is easy to show that if $\rho = 0{\cdot}5$ then ANCOVA with p pre-treatment measurements has the same variance as CHANGE with $p+1$ pre-treatment measurements. For $\rho > 0{\cdot}5$, which is quite likely in practice, this gap between the two methods is narrowed more rapidly.

For instance, Table I compares ANCOVA and CHANGE for $r = 10$ post-treatment visits and $p = 1, \ldots, 5$ pre-treatment visits, all variances being expressed as a proportion of the ANCOVA variance for $p = 1$. For ANCOVA, addition of more pre-treatment visits is always helpful, but especially so if ρ is large. For instance, if $\rho = 0{\cdot}7$, then having a second pre-treatment visit reduces the variance by 36 per cent. Further somewhat less substantial gains are made by adding a third

ssuming

Table I. The dependence of the variances for ANCOVA and CHANGE on the number of pre-treatment measurements p and the equi-correlation ρ between time points assuming $r = 10$ post-treatment visits. For each ρ, variances are divided by the variance for ANCOVA with $p = 1$

ρ	Analysis	Number of pre-treatment measurements, p 1	2	3	4	5
0·3	ANCOVA	1·000	0·827	0·719	0·645	0·591
	CHANGE	2·750	1·500	1·083	0·875	0·750
0·5	ANCOVA	1·000	0·722	0·583	0·500	0·444
	CHANGE	1·833	1·000	0·722	0·583	0·500
0·7	ANCOVA	1·000	0·640	0·490	0·407	0·355
	CHANGE	1·375	0·750	0·542	0·438	0·375
0·9	ANCOVA	1·000	0·574	0·421	0·343	0·296
	CHANGE	1·100	0·600	0·433	0·350	0·300

pre-treatment visit, and so on. This proportionate gain for ANCOVA, as shown in Table I for $r = 10$, is reduced slightly for a smaller number of post-treatment visits.

Table I also reveals that CHANGE becomes closer in efficiency to ANCOVA as the number of pre-treatment readings is increased. For instance, for $p = 5$ and $\rho = 0{\cdot}7$ the variance reduction for ANCOVA is only 5 per cent. This is because the observed pre-treatment mean more closely estimates the true pre-treatment level for each patient. Consequently the 'regression to the mean' problem in a mean changes analysis is reduced and the estimated slope $\hat{\beta}$ in ANCOVA becomes closer to unity.

In some repeated measures designs there may be *a fixed total number of visits* $p + r = t$, and we can therefore only increase the number of pre-treatment visits p at the expense of the number of post-treatment visits r. Then

For CHANGE:

$$\text{variance} = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\sigma^2 \frac{t(1-\rho)}{p(t-p)}$$

For ANCOVA:

$$\text{variance} = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\sigma^2 \frac{t(1-\rho)}{p(t-p)}\left[1 - \frac{(t-p)(1-\rho)}{t[1+(p-1)\rho]}\right].$$

For CHANGE, the variance is minimized for $p = r$, that is equal numbers of pre and post readings when t is even and $p = (t-1)/2$ or $(t+1)/2$ when t is odd.

However, it is more important to consider the choice of p for ANCOVA given a fixed total number of visits $p + r = t$. In general, the 'minimum variance' choice of p for any given t becomes larger as ρ increases, because the pre-treatment readings are of greater use, that is $\hat{\beta}$ becomes larger. More specifically, we can show that for any choice of integer p' then, if $\rho = 1/(t-2p')$, the variances of ANCOVA for $p = p'$ and $p = p' + 1$ are the same. If $\rho < 1/(t-2p')$ then $p = p'$ produces a smaller variance, and if $\rho > 1/(t-2p')$ then $p = p' + 1$ produces a smaller variance. Thus, the optimal choice is $p = p'$ when ρ lies between $1/[t-2(p'-1)]$ and $1/(t-2p')$ for $p' > 0$. Also $p = 0$ if $\rho < 1/t$.

For example, if $t = 10$ measurements in total, we would set $p' = 5$ and divide them equally between pre- and post-treatment readings if $\rho > 1/2$ and set $p' = 4$ if $1/2 > \rho > 1/4$. Smaller values of ρ are unlikely to occur in practice. Hence, if the aim is to minimize the ANCOVA variance, p should be not much smaller than $t/2$, since precision of the individual's pre-treatment mean level is almost as important as precision of the post-treatment mean level.

Of course, reduction in variance is not the only criterion affecting the choice of p. We usually wish to concentrate on the post-treatment readings to describe the shape of mean change over time (for example, is the treatment difference constant, increasing or peaked?) and post-treatment measurements may be required at certain intervals for patient monitoring. Departure from the equi-correlation assumption is also relevant. For instance, if the average correlation between pre and post readings was considerably lower than the average pairwise correlation between pairs of post-treatment readings then the 'minimum variance' p would be further from $t/2$. Nevertheless, the above results appropriately reflect the merit of having multiple pre-treatment readings if practicable.

However, it may sometimes be unfeasible or unethical to obtain multiple pre-treatment measurements at adequate intervals. For instance, if randomization must occur soon after the first visit, there may be no opportunity for repeat pre-treatment visits or their spacing may have to be so close in time that they do not provide sufficiently independent measurements to improve estimation of the subject's 'true baseline'. In most applications, it may be difficult to define what is an adequate minimum spacing, though having $p > 1$ can only do good!

It should be noted that the greatest gain in efficiency is by having $p = 2$ rather than $p = 1$. For instance, with $t = 10$ readings in all and $\rho = 0{\cdot}7$, the reduction in ANCOVA variance for $p = 2$ versus $p = 1$ is 34 per cent while for $p = 5$ versus $p = 1$ the reduction is 53 per cent. In practice, some compromise is needed between precision of overall treatment effect estimation (p sufficiently large) and adequate description of the time pattern of treatment response (r sufficiently large).

The statistical consequences of increasing the number of post-treatment readings r is the same for all three methods of analysis. Under the equi-correlation assumption the reduction in each variance by having $r + 1$ rather than r post-treatment readings is equal to

$$\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\frac{\sigma^2(1-\rho)}{r(r+1)}.$$

The practical consequence of this reduction in variance for increasing r might best be viewed in the context of power calculation, as described in the next section.

6. SAMPLE SIZE DETERMINATION FOR REPEATED MEASURES DESIGN

Returning to the notation of Section 2, we consider the alternative hypothesis $\bar{\mu}_A^{\text{post}} - \bar{\mu}_B^{\text{post}} = \delta$. Also, in the conventional approach to power calculation, we define α and β as the type I and type II errors for ANCOVA.

It is convenient to assume that sample sizes are sufficiently large that the normal approximation to the t-distribution can be applied. In that case, for two equal sized treatment groups of size n we require for general Σ that

$$n = \frac{2\left(\bar{\Sigma}_{\text{post}} - \dfrac{\bar{\Sigma}_{\text{mix}}^2}{\bar{\Sigma}_{\text{pre}}}\right)}{\delta^2} f(\alpha, \beta),$$

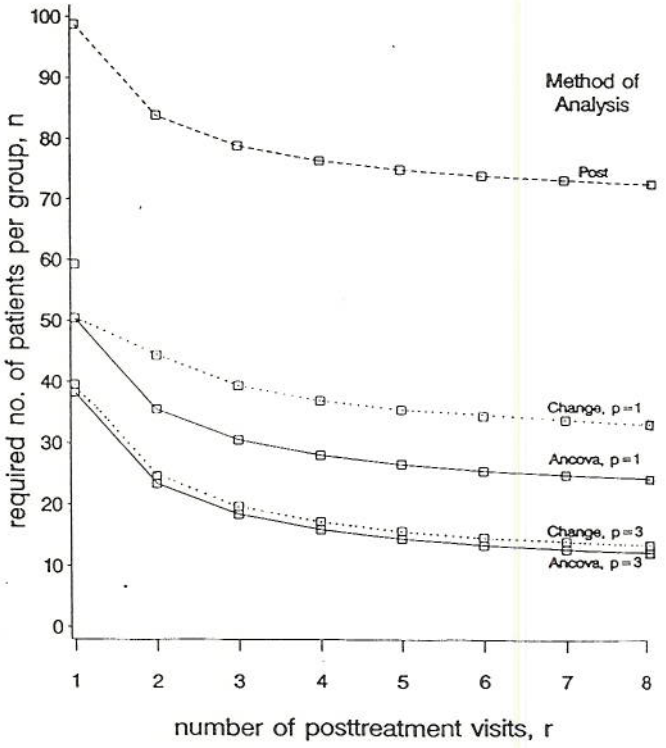

Figure 4. An example of power calculations for a repeated measures design, alternative hypothesis $\delta = 0{\cdot}4\sigma$, $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}2$ (assuming $\rho = 0{\cdot}7$)

where $f(\alpha, \beta) = [z(\alpha/2) + z(\beta)]^2$, $z(x)$ being the standardized normal deviate exceeded with probability x. Under the compound symmetry assumption this becomes

$$n = \frac{2\sigma^2}{\delta^2}\left[\frac{1 + (r-1)\rho}{r} - \frac{p\rho^2}{1 + (p-1)\rho}\right] f(\alpha, \beta).$$

For the other two methods of analysis, POST and CHANGE, we have respectively

$$n = \frac{2\sigma^2}{\delta^2}\left[\frac{1 + (r-1)\rho}{r}\right] f(\alpha, \beta), \quad n = \frac{2\sigma^2}{\delta^2}\left[\frac{1 + (r-1)\rho}{r} - \frac{(p+1)\rho - 1}{p}\right] f(\alpha, \beta).$$

For illustration, consider the alternative hypothesis $\delta = 0{\cdot}4\sigma$, and let $\rho = 0{\cdot}7$, often a realistic value for practical use. Figure 4 shows the required sample size n in each group for a variety of study designs and analysis approaches: for $r = 1, \ldots, 8$ post-treatment measurements, for $p = 1$ or 3 pre-treatment measurements and for POST, CHANGE and ANCOVA.

The simplest possible design has $r = 1$ and $p = 0$. The POST analysis (a two-sample t-test) requires around $n = 100$ patients per group. Increasing the number of post-treatment readings has some effect on decreasing n, but with no use of pre-treatment readings n remains at around 75 even with $r = 8$.

The CHANGE analysis with $p = 1$ pre-treatment measurement (a two-sample t-test comparing mean changes) leads to a required n around 60 for $r = 1$ post-treatment measurement, which can be reduced to $n < 40$ if r is increased to 4 or more post-treatment measurements. The superiority

of ANCOVA is illustrated by a further fall in sample size. For instance, with $p = 1$ and $r \geqslant 4$ we can reduce n to below 30 if ANCOVA is used.

The advantage of increasing the number of pre-treatment measurements is substantial. For instance, with $p = 3$ and $r \geqslant 4$ ANCOVA requires $n < 20$ patients per group. For $p = 3$, CHANGE is similar to ANCOVA.

The more widespread use of such power calculation formulae may add greatly to a sensible choice of n, p and r in repeated measures designs. While the compound symmetry assumption is unlikely to be true, it is often not wildly off the mark and so use of these simple formulae should give an adequate estimate of order of magnitude for n. A plausible value of ρ in the range 0·5 to 0·75 should usually be appropriate. Perhaps the greatest difficulty lies in choosing an appropriate value for δ/σ, but this problem applies to any power calculation with quantitative data expressed as means. Note that σ here is the between-patient standard deviation, although the ANCOVA and CHANGE analyses are within-patient in essence.

7. BIAS IN ESTIMATION IF PRE-TREATMENT MEANS DIFFER

Under the simple model in Section 2 for a randomized clinical trial, $\bar{\mu}_A^{\text{pre}} = \bar{\mu}_B^{\text{pre}}$ and so the expected value of $\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}}$ is zero. Accordingly, at the design stage (before $\bar{x}_{A..}^{\text{pre}}$ and $\bar{x}_{B..}^{\text{pre}}$ are observed), all three methods of analysis produce (on average) unbiased estimates of $\bar{\mu}_A^{\text{post}} - \bar{\mu}_B^{\text{post}}$. However, for any particular observed pre-treatment difference in means $\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}} = d^{\text{pre}} \neq 0$ there exists scope for bias in all three methods.[13]

One rationale behind ANCOVA is that the covariance adjustment removes that component of the observed difference in post-treatment means that is predicted purely on statistical grounds from the observed difference in pre-treatment means. However, as explained by Snedecor and Cochran,[14] this removal of bias due to inequality of pre-treatment means is only true if there is no measurement error in pre-treatment readings.

The concept of measurement error is not easy to define in practice. For example, with measurement of serum cholesterol there is laboratory variation but also short-term within-subject variability over time,[15] both of which count as 'measurement error' around a true underlying serum cholesterol level. While the (technical) laboratory component is often easily determined it is liable to under-represent the true extent of 'measurement error'.

We define a variance σ_e^2 for measurement error, and suppose this is a subcomponent of the overall variance σ^2 defined in Section 2, and that both σ^2 and σ_e^2 are the same for all time points. Then, this measurement error results in an attenuation of the slope β in ANCOVA compared with regression on the true underlying (but unknown) pre-treatment means for each patient. For $p = 1$ pre-treatment readings, the expected value of the observed slope

$$\beta_{\text{obs}} = [(\sigma^2 - \sigma_e^2)/\sigma^2]\, \beta_{\text{true}},$$

as shown by Snedecor and Cochran.[14] Then, for any observed pre-treatment difference $\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}} = d^{\text{pre}}$, the bias in ANCOVA is $(\beta_{\text{true}} - \beta_{\text{obs}})d^{\text{pre}} = [\sigma_e^2/(\sigma^2 - \sigma_e^2)]\beta_{\text{obs}} d^{\text{pre}}$. For $p > 1$ pre-treatment measurements the attenuation in slope becomes less marked, and specifically

$$\beta_{\text{obs}} = \frac{\left(\bar{\Sigma}_{\text{pre}} - \dfrac{\sigma_e^2}{p}\right)}{\bar{\Sigma}_{\text{pre}}} \beta_{\text{true}}.$$

Since $\beta_{\text{obs}} = \bar{\Sigma}_{\text{mix}}/\bar{\Sigma}_{\text{pre}}$, the bias of ANCOVA is given by

$$\frac{\bar{\Sigma}_{\text{mix}}}{\bar{\Sigma}_{\text{pre}}} \frac{\sigma_e^2}{(p\bar{\Sigma}_{\text{pre}} - \sigma_e^2)} d^{\text{pre}}.$$

POST is equivalent to ANCOVA with β forced equal to zero, and hence the bias of POST is

$$\beta_{\text{true}} d^{\text{pre}} = \frac{\bar{\Sigma}_{\text{mix}}}{\bar{\Sigma}_{\text{pre}} - \frac{\sigma_e^2}{p}} d^{\text{pre}}.$$

CHANGE is equivalent to ANCOVA with β forced equal to one, and hence tends to overcorrect for any imbalance in pre-treatment means. Specifically, the consequent bias of CHANGE is

$$-(1 - \beta_{\text{true}}) d^{\text{pre}} = -\left[1 - \frac{\bar{\Sigma}_{\text{mix}}}{\bar{\Sigma}_{\text{pre}} - \frac{\sigma_e}{p}}\right] d^{\text{pre}}.$$

In each case the bias is $c\, d^{\text{pre}}$ and so, assuming known covariance matrices, and under simple randomization,

$$\text{expected}(\text{bias}^2) = c^2 \exp(d^{\text{pre}})^2 = c^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right) \bar{\Sigma}_{\text{pre}}.$$

We could then consider for each analysis the mean squared error = sampling variance + expected (bias^2), the first component having previously been defined and explored in Sections 2 to 5. However, this 'mean squared error' approach has certain problems: the assumption of simple unstratified randomization may not apply, and knowledge of σ_e^2 is usually lacking. Also, while 'mean squared error' may be conceptually useful when planning a repeated measures study, that is prior to knowing the observed $\bar{x}_{A..}^{\text{pre}} - \bar{x}_{B..}^{\text{pre}}$, once the data are observed it may be better to consider the actual bias of each method of analysis separately from its sampling variance.

It is useful to quantify the extent of bias in practice for the three approaches. First, we would hope that measurement error accounts for a small proportion of the observed between-patient variance at each time point since otherwise we might question the outcome measure's suitability. Thus, in most practical circumstances the bias of ANCOVA should be small. Having more than one pre-treatment measurement will reduce this bias further, since it is approximately proportional to $1/p$. For POST, if we assume σ_e^2 is negligibly small, and adopt the compound symmetry assumption, then the bias $= \{p\rho/[1 + (p-1)\rho]\} d^{\text{pre}}$. Under the same assumption, for CHANGE the bias $= \{-(1-\rho)/[1 + (p-1)\rho]\} d^{\text{pre}}$. For $p = 1$, this means that the POST bias $= \rho d^{\text{pre}}$ and the CHANGE bias $= -(1-\rho) d^{\text{pre}}$. For $\rho > 0{\cdot}5$ (which is usually true), POST contains more bias than CHANGE. Furthermore, for $p > 1$ pre-treatment measurements, this aspect of inferiority for POST becomes more marked. For instance, if $p = 3$ and $\rho = 0{\cdot}7$ (say), then the POST bias $= 0{\cdot}875\, d^{\text{pre}}$ while the CHANGE bias $= 0{\cdot}125\, d^{\text{pre}}$. However, with more pre-treatment readings we can expect d^{pre} to become smaller.

Overall, if there exists a pre-treatment difference, then POST may be seriously biased. CHANGE may also contain a degree of bias, especially if the correlations between pre and post measurements are relatively small, but this bias will be reduced considerably if the number of pre-treatment measurements is increased.

8. SOME EXAMPLES

We now illustrate the practical relevance of the above design and analysis issues for widespread use in clinical trials with repeated measurements. First, we return to analysis of the example presented in Section 2 and Figure 1, and then the correlation structures of several other repeated

Table II. ANCOVA, CHANGE and POST analyses for the CPK data. $n = 76$ patients in each treatment group, $r = 8$ post-treatment measurements, $p = 1$ or 3 pre-treatment measurements; $\hat{\beta}$ is estimated regression coefficient

	Number of pre-treatment measurements	Estimated mean difference in CPK (IU/l)	Standard error (IU/l)	t-statistic	P
ANCOVA ($\hat{\beta} = 0{\cdot}83$)	3	− 0·066	0·021	3·24	0·001
ANCOVA ($\hat{\beta} = 0{\cdot}63$)	1	− 0·043	0·025	1·72	0·09
CHANGE	3	− 0·062	0·022	2·89	0·004
CHANGE	1	− 0·023	0·030	0·77	0·44
POST		− 0·085	0·037	2·31	0·02

measures trials are summarized to provide practical recommendations on the degree and pattern of correlation we might commonly expect.

From the discussion above, the most appropriate method of analysis for the data in Figure 1 is ANCOVA based on each patient's mean of the eight post-treatment measurements with the mean of the patient's three pre-treatment measurements as covariate. Table II shows CHANGE and POST for comparison, and also includes for illustration ANCOVA and CHANGE as if only a single pre-treatment measurement (visit 0) had been available.

ANCOVA is seen to produce a smaller standard error and hence stronger evidence of a treatment difference, especially if the mean of all three baseline readings is used as a covariate. Since $\hat{\beta}$ is close to 1 in this case the CHANGE analysis is only marginally inferior. POST suffers from two problems: the standard error is much larger, and also failure to take account of the slightly higher average pre-treatment mean level on active drug leaves an upward bias in the estimated treatment effect. With just a single pre-treatment reading (visit 0) rather than the mean of three, the standard errors for ANCOVA and CHANGE are substantially increased. Given the more pronounced pre-treatment imbalance at visit 0, the CHANGE analysis is prone to a downward bias, this being related to the smaller $\hat{\beta}$ for ANCOVA when $p = 1$.

To explore more generally the correlation structures of repeated measures in clinical trials, a number of examples are summarized in Table III. These examples represent the most recent experience of such trials that we and our colleagues in the Medical Statistics Unit have encountered and all have two randomized treatment groups. The aim is to obtain a reasonably representative sample of trials covering a variety of diseases and quantitative outcome measures.

For each trial Table III lists the disease, the number of randomized patients, the numbers of pre- and post-treatment measurements and the mean time between post-treatment measurements, and then for each outcome measure the mean correlations for pre–pre, pre–post (mixed) and post–post pairs of time points and the estimated slope (decrease) in correlation with 'time' between visits (where 'time' denotes the number of visits apart). In nearly all instances the post-treatment visits were at equally spaced intervals.

Certain general characteristics emerge from these trials. The correlations between post-treatment visits mostly average between 0·6 and 0·8. A similar magnitude of correlation exists between pre-treatment visits, when $p \geqslant 2$. The average mixed pre–post correlation is mostly of similar magnitude, but with a tendency to be slightly lower. Most examples show a slight decline in correlation (amongst post-treatment visits) as the time interval between measurements increases.

It is interesting to observe one or two exceptions to this general pattern. The hypertension trial in elderly patients had somewhat lower correlations for blood pressure, and this can be attributed

Table III. Summary of the correlations in repeated measurements from a sample of clinical trials

Disease	Number of patients	Number of visits pre (p)	Number of visits post (r)	Mean time between post visits (months)	Outcome measures	Mean correlation pre	Mean correlation mixed	Mean correlation post	Estimated slope
Coronary heart disease	152	3	8	1·5	CPK	0·65	0·62	0·67	− 0·012
					ALAT	0·69	0·64	0·67	− 0·017
					ASAT	0·69	0·70	0·76	− 0·006
					Alkaline phosphatase	0·79	0·73	0·75	0·004
Coronary heart diesease	219	2	4	3	HDL	0·74	0·74	0·84	− 0·006
					Triglycerides	0·68	0·56	0·56	− 0·066
					Total cholesterol	0·65	0·52	0·65	− 0·011
Hypertension	55	3	12	1	Heart rate	0·64	0·56	0·61	− 0·010
					Systolic blood pressure	0·62	0·56	0·70	− 0·006
Hypertension in elderly	3450	1	7	2	SBP	—	0·23	0·44	− 0·029
					DBP	—	0·44	0·55	− 0·024
Intermittent claudication	504	2	2	6	Ankel/arm ratio of systolic blood pressure	0·74	0·62	0·65	—
Angina	251	1	3	4	Treadmill test distance	—	0·53	0·77	− 0·002
Childhood asthma	138	1	10	3	FEV_1	—	0·70	0·81	− 0·006
		1	5	6	PC_{20} (histamine responsiveness)	—	0·47	0·75	− 0·032
Multiple sclerosis	162	1	3	1	Muscle tone score	—	0·70	0·80	− 0·010
Low back pain	459	2	3	8	Back pain score	0·85	0·29	0·75	0·000
HIV infection	545	1	6	4	CD_4 cell count	—	0·68	0·77	− 0·021

* Estimated (by least squares) decrease in correlation per visit apart among the post-treatment visits.

to the fact that treatment regimens were adjusted over time in each patient according to observed blood pressure; for example, a patient whose blood pressure stayed high received additional dosage or supplementary drugs. This is perhaps an unusual adaptive feature not commonly encountered in studies with repeated measures. The low back pain study had a low mixed correlation, and this reflects the fact that a proportion of patients were cured (back pain score = 0) and such prospect of cure was not closely associated with the original severity of disease.

In most of these examples correlations tend to decline slightly over time and mixed correlations are slightly lower, but there is no major departure from the compound symmetry assumptions used earlier. In fact, we could suggest that in the absence of prior knowledge on correlation structure at the design stage of a repeated measure trial, the assumption of compound symmetry with $\rho = 0{\cdot}65$ (say) would be unlikely to lead to major errors of judgement in evaluating choices of p, r and n. However we caution that such a statement requires detailed investigation under alternative correlation structures.

9. THE EFFECT OF UNEQUAL CORRELATIONS

Although the examples in Section 8 are reassuring regarding the lack of serious departure from compound symmetry, it is nevertheless important to consider the impact that such departure would have on the design recommendations of earlier sections. We suspect non-equal correlations is a more serious problem than inequality of variances, and will focus on alterations to the power calculations (especially Figure 4) as a means of illustrating the design implications of unequal correlations.

Let $\bar{\rho}_{\text{post}}$, $\bar{\rho}_{\text{mix}}$ and $\bar{\rho}_{\text{pre}}$ be the mean pairwise correlations in the post–post, pre–post and pre–pre covariance submatrices Σ_{post}, Σ_{mix} and Σ_{pre} respectively. Then it is easily shown that the variances of treatment differences and the required sample sizes are proportional to the following:

POST:

$$\frac{1 + (r-1)\bar{\rho}_{\text{post}}}{r}$$

CHANGE:

$$\frac{1 + (r-1)\bar{\rho}_{\text{post}}}{r} + \frac{1 + (p-1)\bar{\rho}_{\text{pre}}}{p} - 2\bar{\rho}_{\text{mix}}$$

ANCOVA:

$$\frac{1 + (r-1)\bar{\rho}_{\text{post}}}{r} - \frac{\bar{\rho}_{\text{mix}}^2}{\left(\dfrac{1 + (p-1)\bar{\rho}_{\text{pre}}}{p}\right)}.$$

Therefore, determination of trial size and its dependence on r, p and the method of analysis can all be documented if one knows the values of the three parameters $\bar{\rho}_{\text{post}}$, $\bar{\rho}_{\text{mix}}$ and $\bar{\rho}_{\text{pre}}$. Given the theoretically infinite variety of correlation structures that could exist, one cannot reach completely generalizable quantitative conclusions on these design issues. However, in the light of the above examples we will attempt to elucidate some practical suggestions based on certain realistic departures from compound symmetry.

First, consider $p = 1$ pre-treatment reading and the consequence of having $\bar{\rho}_{\text{mix}}$ different from $\bar{\rho}_{\text{post}}$ ($\bar{\rho}_{\text{pre}}$ is non-existent if $p = 1$). Suppose non-equality of correlations can be represented by

a decline in ρ of magnitude b per visit apart, all visits being equally spaced. If data exist from a previous trial, this slope b can be estimated from the full correlation matrix, as in Figure 2. Then it can be shown that $\bar{\rho}_{post} - \bar{\rho}_{mix} = b(r+1)/6$. For the examples in Table III, the estimated differences $\hat{\bar{\rho}}_{post} - \hat{\bar{\rho}}_{mix}$ are all positive, and although relatively small (median estimated difference across all examples is 0·10) they are mostly greater than the above formula would suggest based on the small estimated slopes in Table III. If power calculations take account of $\bar{\rho}_{mix}$ being less than $\bar{\rho}_{post}$, the sample size reductions in CHANGE and ANCOVA compared with POST become less marked. For instance, reworking the example in Figure 4 with $\bar{\rho}_{mix} = 0{\cdot}6$ and $\bar{\rho}_{post} = 0{\cdot}7$ then for $p = 1$ and any r, ANCOVA has the required n increased by 12·8 and CHANGE has n increased by 19·8.

Next, consider the decline in sample size with increasing r and how this could be affected by unequal correlations. For POST this depends solely on the relationship between $\bar{\rho}_{post}$ and r. Suppose correlations get weaker the further apart this visits are, as often found in our examples. For a fixed total follow-up time T it can be shown that for r equally spaced visits the mean of all pairwise distances is $(r+1)T/3r$. This declines with r (by a maximum of one-third for $r = \infty$ compared with $r = 2$) so that $\bar{\rho}_{post}$ increases with r. Hence, for POST the trends in declining sample size with increasing r in Figure 4 will level off slightly more quickly. In practice, for data with only slightly unequal correlations as in Figure 2, this effect would be negligible. For CHANGE and ANCOVA, we note that increasing r is liable to increase slightly both $\bar{\rho}_{post}$ and $\bar{\rho}_{mix}$, two opposite effects on the trends in sample size with r which will tend to cancel one another out, leaving the equi-correlation declines in required n as reasonably reliable.

When considering the merit of $p > 1$ baseline readings, the extent to which $\bar{\rho}_{pre}$ and $\bar{\rho}_{mix}$ differ from $\bar{\rho}_{post}$ has some bearing on the power calculations. If the repeat baselines are close together $\bar{\rho}_{pre}$ might be increased, whereas having baselines further back in time might reduce $\bar{\rho}_{mix}$, either of these possibilities leading to an increase in the required sample size for ANCOVA. For instance, for $p = 3$ baselines in the Figure 4 example suppose $\bar{\rho}_{pre} = 0{\cdot}8$ and $\bar{\rho}_{mix} = 0{\cdot}6$ while $\bar{\rho}_{post} = 0{\cdot}7$. Then for ANCOVA the required n is increased by 19·5 compared with equi-correlation $\rho = 0{\cdot}7$ (for any value of r). The consequent reduction in sample size for ANCOVA with $p = 3$ instead of $p = 1$ becomes only about half that seen in Figure 4.

Overall, substantial improvements in statistical efficiency with repeat baselines are possible provided $\bar{\rho}_{mix}$ is not radically reduced and $\bar{\rho}_{pre}$ is not too large. The magnitude of benefit is dependent on $\bar{\rho}_{mix}$ and $\bar{\rho}_{pre}$, but like other parameters in power calculation their values may not be known in advance. Thus, while the recommendation to have more than one baseline if possible is of general relevance to repeat measures trials, the precise extent of statistical improvement cannot be reliably predetermined unless one has some prior knowledge (for example from a previous trial) regarding the correlation structure.

10. CONCLUDING REMARKS

The analysis of repeated measures has stimulated a wide range of methodological developments in recent years. However, we feel that a considerable gap has opened up between quite complex statistical theory and the day-to-day reality of statistical reporting in medical journals.

The aim of this paper has been to explore the statistical properties of some simple approaches to repeated measures using summary statistics. While there are many possible summary statistics, we have focused on the mean post-treatment response of each patient as being a logical choice in many such trials. Consequently, ANCOVA using the mean pre-treatment level as a covariate is the preferred method of analysis. In practice, we suspect ANCOVA is not used nearly enough, so

that too many trial reports of quantitative outcome variables, with or without repeated measurements, rely on inferior analyses using just post-treatment values or post–pre differences.

In some trials, a gradual widening of treatment differences over time may be anticipated, for example the effect of drug treatment on CD_4 cell counts in HIV infection, and then the estimated slope over time may be a preferred summary statistic.[16,17] Further evaluation of the relative merits of the slope and the mean response would be of value, as would consideration of the spacing[18] and number of repeated measures in trials with increasing treatment effects over time.

In many repeated measures trials the analysis is complicated by missing values and patient withdrawals. With a summary statistic approach it is relatively easy to define sensible criteria for coping with occasional missing values. However, patient withdrawals provoke greater problems since they are often associated with informative censoring.[19] While a summary statistic approach may be able to use substitute summary measures for individuals with fewer measurements prior to withdrawal, it needs to be recognized that the presence of substantial withdrawals may lead to bias whichever analysis technique is employed.

Finally, we feel that little attention has been given to the statistical design of clinical trials with repeated measurements. Our methods for determining sample size and the number of pre- and post-treatment measurements should be of practical use in the planning of such trials.

We feel that the examples presented support the use of the compound symmetry assumption as a realistic guide to the quantitative planning of clinical trials with repeated measures. It should also be noted that the proposed analyses using summary statistics do not depend on any such assumption. However, cautious recognition of the effects of non-equal correlations on design considerations should be made when necessary, as illustrated in Section 9. Indeed, we would encourage further research into the impact of correlation mis-specification, possibly using a variety of alternative structures, such as autoregression and moving averages. However, we would plead that further methodological developments should keep in touch with the needs of statistical and clinical practitioners who undertake repeated measures trials.

ACKNOWLEDGEMENTS

We are grateful to Simon Thompson, Stephen Evans and the referees for their helpful advice.

REFERENCES

1. Crowder, M. J. and Hand, D. J. *Analysis of Repeated Measures*, Chapman and Hall, London, 1990.
2. Goldstein, H. *Multilevel Models*, Griffin, Oxford, 1987.
3. Milliken, G. A. 'Analysis of repeated measures designs', in D. A. Berry (ed.), *Statistical Methodology in the Pharmaceutical Sciences*, Dekker, New York, 1990.
4. DeKlerk, N. H. 'Repeated warnings re repeated measures', *Australian and New Zealand Journal of Medicine*, **16**, 637–638 (1986).
5. Finney, D. J. 'Repeated measurements: what is measured and what repeats?', *Statistics in Medicine*, **9**, 639–644 (1990).
6. Pocock, S. J., Hughes, M. D. and Lee, R. L. 'Statistical problems in the reporting of clinical trials: a survey of three major medical journals', *New England Journal of Medicine*, **317**, 426–432 (1987).
7. Matthews, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P. 'Analysis of serial measurements in medical research', *British Medical Journal*, **300**, 230–235 (1989).
8. Salsburg, D. C. 'Development of statistical analysis for single dose bronchodilators', *Controlled Clinical Trials*, **2**, 305–317 (1981).
9. O'Brien, P. C. 'Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079–1087, (1984).
10. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. 'The analysis of multiple endpoints in clincial trials', *Biometrics*, **43**, 487–498 (1987).
11. Fleiss, J. L. *Design and Analysis of Clinical Experiments*, Wiley, New York, 1986, Chapter 7.

12. Rouanet, H. and Lepine, D. 'Comparison between treatments in a repeated-measurements design: ano and multivariate methods', *British Journal of Mathematics and Statistical Psychology*, **23**, 147–1 (1970).
13. Blomqvist, N. 'On the relation between change and initial value', *Journal of the American Statistic Association*, **72**, 746–749 (1977).
14. Snedecor, G. W. and Cochran, W. G. *Statistical Methods*, Iowa State Press, 1989, Chapter 14.
15. Thompson, S. G. and Pocock, S. J. 'The instability of serum cholesterol measurements: implications fo screening and monitoring', *Journal of Clinical Epidemiology*, **43**, 783–789 (1990).
16. Laird, N. M. and Wang, F. 'Estimating rates of change in randomized clinical trials', *Controlled Clinica Trials*, **11**, 405–419 (1990).
17. Dawson, J. D. and Lagakos, S. W. 'Analyzing laboratory marker changes in AIDS clinical trials', *Journal of AIDS*, **4**, 667–676 (1991).
18. Morrison, D. F. 'The optimal spacing of repeated measurements', *Biometrics*, **26**, 281–290 (1970).
19. Wu, M. and Bailey, K. R. 'Estimation and comparison of changes in the presence of informative right censoring: conditional linear model', *Biometrics*, **45**, 939–955 (1989).