STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4334–4344
Published online 21 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2682

Change from baseline and analysis of covariance revisited

Stephen Senn*, †

Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow G12 8QQ, U.K.

SUMMARY

The case for preferring analysis of covariance (ANCOVA) to the simple analysis of change scores (SACS) has often been made. Nevertheless, claims continue to be made that analysis of covariance is biased if the groups are not equal at baseline. If the required equality were in expectation only, this would permit the use of ANCOVA in randomized clinical trials but not in observational studies. The discussion is related to Lord's paradox. In this note, it is shown, however that it is not a necessary condition for groups to be equal at baseline, not even in expectation, for ANCOVA to provide unbiased estimates of treatment effects. It is also shown that although many situations can be envisaged where ANCOVA is biased it is very difficult to imagine circumstances under which SACS would then be unbiased and a causal interpretation could be made. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: change score; analysis of covariance; Lord's paradox; repeated measures; baselines

1. INTRODUCTION

The case for using analysis of covariance as a means of adjusting for baseline measurements has been made many times in the statistical literature, including in *Statistics in Medicine*. Some authors have stressed its ability to deal with conditional bias [1, 2]. Others have extensively investigated its positive effect on efficiency [3, 4]. Nevertheless, some authors continue to questions its appropriateness, claiming, for example that measurement error implies that analysis of covariance does not produce conditionally unbiased estimates [5]. However, as has been pointed out, this is based on a failure to appreciate that in a randomized clinical trial it is observed imbalance that needs to be dealt with and therefore conditioning on observed covariates forms a valid approach to adjustment [6–9].

However, other objections are sometimes raised. For example, a paper by Liang and Zeger in *Sankyha* examines in detail various strategies for estimating treatment effects in repeated measures

*Correspondence to: Stephen Senn, Department of Statistics, 15 University Gardens, University of Glasgow, Glasgow G12 8QQ, U.K.
†E-mail: stephen@stats.gla.ac.uk

Received 16 June 2006
Accepted 6 July 2006

designs [10]. The simplest case considered by them corresponds to a design with two treatments (say an experimental treatment and a standard control) and two measurements: at baseline prior to treatment and at outcome. In connection with that design they make the claim (on p. 138) that except where the expected difference in baseline values between treatment groups is zero, then analysis of covariance (ANCOVA) conditioning on the baseline values gives a biased estimate of the treatment effect. On the other hand, they claim that a straightforward analysis of differences between outcome and baselines (such differences are sometimes referred to as *change scores*) will produce an unbiased, if less efficient, estimate of the treatment effect.

They are not the first to have made these claims. In 1986, in an interesting discussion of ANCOVA *versus* the change score approach, Samuels in *Controlled Clinical Trials* made a similar claim in comparing ANCOVA to the t-test based on the change-score [11]. She stated, '... the choice between the t-test and ANCOVA depends on whether one is willing to assume that the following condition holds: $\mu_1 = \mu_2$.' (p. 327). Here the symbols μ_1, μ_2 represent the expected values at baseline in the two treatment groups.

In this note we show that these claims are not correct. It is *not* a necessary condition for the unbiasedness of ANCOVA for the expected values at baseline between the two groups to be equal. Furthermore, there are some circumstances where ANCOVA is unconditionally unbiased but the simple analysis of change scores (SACS) is not. The use of the word *simple* here in not meant to be pejorative. The qualifier is used to distinguish such an analysis from more complex approaches in which the change score is used but covariates can also be fitted. Of course, if the baseline is such a covariate, then the result is formally equivalent to ANCOVA [12].

The outline of this note is as follows. In Section 2 we present a simple counter-example to show a case where a trial with unbalanced means at baselines may be validly analysed by ANCOVA but not validly analysed by SACS. In Section 3 we show the general conditions for ANCOVA to be valid. In Section 4 we present a general model for baselines and outcomes and illustrate the origin of the claim that ANCOVA is biased except where expectations at baseline are equal. We then proceed in Section 5 to examine if circumstances can exist in which SACS is unbiased and ANCOVA is not. Finally in Section 6 we discuss the implications of our results.

We deliberately do not start by presenting a model for repeated measures data. There are two reasons. First, an approach based on specifying a given model will yield valid conclusions if the model is correct but does not address directly the validity of the model. It can provide sufficient conditions for a conclusion to be correct but not necessary ones. The correct approach to find the necessary conditions is to state the conclusions and see what they imply. Second, it is in fact our claim that many conventionally used models for repeated measures are potentially misleading. What the origin of this potential confusion is will become clearer once the validity or otherwise of ANCOVA and SACS have been examined.

First, however, we provide a brief justification for yet another paper on this topic. The debate is not new. The fact that SACS and ANCOVA can lead to surprisingly different conclusions was presented by Lord nearly 40 years ago in a paper [13] that outlines the paradox that now bears his name. He described a situation in which two groups of individuals were measured at baseline and at outcome. In his original example weight was the outcome variable and the groups were university students visiting two dining halls. The groups were weighed in September and in the following June, mean weight being quite different for the two halls at outset. In neither group did the mean change over time. In consequence SACS simply compared a difference of zero from each group and returned a measured 'effect' of zero. However, whatever difference there was at baseline was

DOI: 10.1002/sim

the difference observed at outcome and since the correlation was not perfect, ANCOVA subtracted a fraction (assuming unchanging variability) of the difference at baseline from that at outcome. Hence a non-zero (and statistically significant) 'effect' was returned. This was the paradox: no change observed over time in either hall but analysis of covariance pointing to a difference. More generally the paradox is simply that whereas under any circumstances where differences at baseline may be expected to be zero (as for example over all randomizations in a clinical trial) SACS and ANCOVA may be expected (in the statistical sense) to give the same answer, where this is not the case, and groups may be expected to differ at baseline, they will not be expected to deliver the same answer, raising the question as to which is 'correct'.

This paradox has been ably analysed by Holland and Rubin [14] in 1983 and subsequently by Wainer [15] in 1991 and more recently in 2004 by Wainer and Brown [16], who demonstrated that the validity of one or other approach depended on un-testable assumptions if the purpose of analysis was to answer a causal question. It might be thought that this paradox is not in need of further analysis. However, it will be maintained here that the issue has life in it yet. For example, Wainer's position seems to have evolved between 1991 and 2004 to a rather more positive view of ANCOVA and it will be shown below by a simple but compelling counterexample that the claim of Liang and Zeger [10], that ANCOVA is never justified in the repeated measures context unless baseline means are equal between groups is not correct. Furthermore, the paper finishes by offering a challenge to any reader who is not convinced by its arguments to design an experiment in which SACS would give a valid answer, ANCOVA would not and a causal interpretation could be given.

2. A SIMPLE COUNTER EXAMPLE

We suppose a trial in hypertension in which there is a single measurement of diastolic blood pressure (DBP) at baseline made to a precision of 1 mmHg. We suppose that if and when patients provide consent to enter the trial, a decision is made as to whether they will in fact be entered based on their DBP. Only patients with a DBP to the nearest mm of Hg of either X_1 or X_2, where $X_1 < X_2$ are two integers, will be entered into the trial. This procedure is, of course, possibly unethical and certainly extremely stupid. However its stupidity does not end there. Having been screened as suitable, patients are randomized in the ratio $r : (1-r)$, $0.5 < r < 1$ to the active treatment, A, compared to the control, C, if DBP $= X_1$ mmHg and in the ratio $(1-r) : r$ if DBP $= X_2$ mmHg. Recruitment does not end until n patients have been recruited from each stratum. (To take a concrete example suppose $X_1 = 95$, $X_2 = 105$, and $r = \frac{2}{3}$.) This, of course, compounds the inefficiency of only selecting sub-groups of eligible patients with both an inefficient allocation and an inefficient stopping rule for recruitment. We now add an element of further fantasy to this example by assuming that there are no drop-outs from the trial, in fact no missing measurements at all and the patients are all completely compliant. At the end of the trial the allocation of patients is as given in Table I.

Clearly the trial is one no investigator would ever run. However, a single counter-example, however far-fetched, is sufficient to prove the falsity of a general claim and in fact, once this counter-example has been presented, we shall proceed to show that for a much wider class of design the same claims apply. Furthermore, this example involves stratification on a binary covariate (if a rather unusual one) and that is not only extremely common but instructive for understanding ANCOVA.

DOI: 10.1002/sim

Table I. Allocation of patients in the trial of hypertension.

		Stratum		
		$s=1$ X_1 mmHg	$s=2$ X_2 mmHg	Total
Treatment group	Active	nr	$n(1-r)$	n
	Control	$n(1-r)$	nr	n
	Total	n	n	$2n$

Clearly, within each stratum, s, an unbiased estimate $\hat{\tau}_s$ of the treatment effect for that stratum may be constructed by comparing the two means at outcome. We thus have

$$\hat{\tau}_s = \overline{Y}_{s,A} - \overline{Y}_{s,C}, \quad s = 1, 2$$

$\overline{Y}_{s,A}$ is the stratum mean at outcome under active treatment and $\bar{Y}_{s,C}$ is the mean under control.

Now suppose we wish to construct a general summary of the treatment effect over both strata. Under a standard assumption of homoscedasticity then we have that the treatment estimates $\hat{\tau}_1, \hat{\tau}_2$ are measured with equal precision. In the absence of any prior information and provided no attempt is to be made to recover any inter-stratum information, there is thus no reason not to use the simple average of these treatment estimates as the estimate of the average treatment effect, which we may calculate as

$$\hat{\tau} = \frac{(\hat{\tau}_1 + \hat{\tau}_2)}{2} = \frac{[(\overline{Y}_{1,A} - \overline{Y}_{1,C}) + (\overline{Y}_{2,A} - \overline{Y}_{2,C})]}{2} \tag{1}$$

In any case, whether or not (1) is accepted as being the estimator of choice, it is clear that

$$E[\hat{\tau}] = \frac{(\tau_1 + \tau_2)}{2}$$

so that $\hat{\tau}$ *is* an unbiased estimator of the average treatment effect and, in fact, under an assumption of additivity, which implies $\tau_1 = \tau_2 = \tau$, is an unbiased estimate of the treatment effect for any patient in the trial. However, a little thought shows that this estimator *is* the ANCOVA estimator. This can be seen by considering that stratification on stratum membership is equivalent to carrying out a regression using a dummy variable indicator. One possible coding of this dummy variable is X_1, if $s=1$, X_2, if $s=2$ and this is obviously the same as ANCOVA using the baseline values. (There would be a slight difference as regards estimation of the residual variance since stratification leaves one fewer degree of freedom for estimating error than analysis of covariance.)

Note that the average responses within treatment groups across strata may be expressed as

$$\overline{Y}_A = (1-r)\overline{Y}_{1,A} + r\overline{Y}_{2,A}, \overline{Y}_C = r\overline{Y}_{1,C} + (1-r)\overline{Y}_{2,C}$$

and hence the naïve difference at outcome is

$$\hat{\tau}' = \overline{Y}_A - \overline{Y}_C = [(1-r)\overline{Y}_{1,A} + r\overline{Y}_{2,A}] - [r\overline{Y}_{1,C} + (1-r)\overline{Y}_{2,C}] \tag{2}$$

DOI: 10.1002/sim

Note that if we write $\overline{X}_A, \overline{X}_C$ for the means at baseline in the two treatment groups averaged over both strata, then we have $\overline{X}_A = rX_1 + (1-r)X_2$, $\overline{X}_C = (1-r)X_1 + rX_2$. Hence, the difference at baseline between the two groups is

$$\overline{X}_A - \overline{X}_C = (2r-1)(X_1 - X_2) \tag{3}$$

Hence, as is well known, the ANCOVA adjusted estimate, $\hat{\tau}$, is the difference at outcome minus the difference at baseline weighted by the regression of outcome on baseline, $\hat{\beta}$, we have

$$\hat{\tau} = \hat{\tau}' - \hat{\beta}(2r-1)(X_1 - X_2) \tag{4}$$

In fact, because our example is degenerate in that there are only two possible values of X and $\hat{\beta}$ has the particularly simple form

$$\hat{\beta} = \frac{1}{2}\left[\frac{\overline{Y}_{1,A} - \overline{Y}_{2,A}}{X_1 - X_2} + \frac{\overline{Y}_{1,C} - \overline{Y}_{2,C}}{X_1 - X_2}\right] = \frac{(\overline{Y}_{1,A} + \overline{Y}_{1,C}) - (\overline{Y}_{2,A} + \overline{Y}_{2,C})}{2(X_1 - X_2)} \tag{5}$$

Note that $\hat{\beta}$ as given in (5) is simply the difference at outcome between strata averaged over both treatments groups divided by the associated change in baseline. It also follows from the fact that (3) is the mean difference between treatment groups at baseline that the SACS estimator is given by

$$\hat{\tau}^* = \hat{\tau}' - (2r-1)(X_1 - X_2) \tag{6}$$

Note that since (4) = (1) and (1) is unbiased by construction and since in general $\hat{\beta} \neq 1$, then $E[\hat{\tau}^*] \neq E[\hat{\tau}]$ and it follows that whereas the ANCOVA estimator is unbiased the SACS estimator is biased. We have thus constructed a counterexample to the claims of Liang and Zeger [10] and Samuels [11].

3. NECESSARY CONDITIONS FOR THE UNBIASEDNESS OF ANCOVA

We now proceed to consider general conditions for the unbiasedness of the ANCOVA estimate. We simply assume in this section that we have a single continuous covariate, X, at baseline. Typically in the philosophy of repeated measures designs this is assumed to be a measurement of the same kind as that at outcome. Note this assumption is necessary for SACS to produce a meaningful estimate since otherwise it will not satisfy dimensional analysis. On the other hand, neither ANCOVA nor (trivially) a simple analysis of outcomes needs the assumption to produce an answer that is expressed in meaningful units.

Let ΔY be the observed difference in means at outcome between two treatments groups and ΔX be the corresponding covariate difference at baseline. Then we have the following three estimators [17]:

$$\text{Unadjusted:} \quad \hat{\tau}' = \Delta Y \tag{7}$$

$$\text{SACS:} \quad \hat{\tau}^* = \Delta Y - \Delta X \tag{8}$$

$$\text{ANCOVA:} \quad \hat{\tau} = \Delta Y - \hat{\beta}\Delta X \tag{9}$$

DOI: 10.1002/sim

Now, we suppose that $E[\Delta Y] = \tau + B_Y$. Here B_Y is simply some bias. If $B_Y = 0$ then $\hat{\tau}'$ is unbiased for the treatment effect. We also suppose that $E[\Delta X] = B_X$. Note that the use of expectation operators here might be regarded as somewhat glib. It will depend very much on circumstances as to what 'universe' one considers has been averaged in expectation. For example in a trial involving an element of randomization $E[\Delta X]$ might be the average mean baseline difference over all possible allocations. If one is interested in conditional inference it might (trivially) be the observed mean difference for the trial actually run. $E[\Delta Y]$, on the other hand, may be more difficult to define. Some at this stage might wish to introduce heavy counterfactual machinery of the sort used to considerable effect by Rubin [18]. Others may argue that such counterfactuals are misleading [19]. We will duck the issue here apart from noting that there may be some circumstances in which no meaning could be given to these terms and that then one is perhaps best avoiding talking about causality at all. This issue becomes relevant in Section 5 and is picked up there.

Assuming for the moment, however, that some meaning can be given to these terms, then if $B_X = B_Y$ then $\hat{\tau}^*$ is unbiased for τ. This latter assumption does, indeed, follow from a standard repeated measures model but is not in general necessary and we avoid making it here. We shall return to this point. To establish the bias of (9) we proceed as follows:

$$\begin{aligned} E[\hat{\tau}] &= E[\Delta Y] - E[\hat{\beta}\Delta X] \\ &= \tau + B_Y - E[\hat{\beta}]E[\Delta X] \\ &= \tau + B_Y - \beta B_X \end{aligned}$$

(Note that the step in the middle is justified by the independence of $\hat{\beta}$ and ΔX.) Hence we have that the bias of $\hat{\tau}$ is given by

$$E[\hat{\tau}] - \tau = B_Y - \beta B_X \tag{10}$$

It thus follows that ANCOVA is unbiased provided that

$$\beta = \frac{B_Y}{B_X} \tag{11}$$

Now, in the repeated measures context, if we make the often made (but not necessarily reasonable) assumption that the variance at baseline is the same as the variance at outcome we have that β is simply the correlation coefficient, ρ, between the two. Hence we have in practice $\beta < 1$ so that if

$$B_Y = B_X \tag{12}$$

then ANCOVA is indeed biased and SACS is unbiased. As regards this assumption we shall show in the next section that it is indeed implicit in many of the standard models used in a repeated measures context but that it is neither necessary nor even necessarily reasonable. In a further section we will show that, indeed, there are a wide variety of circumstances under which (11) will be true and (12) will not be true. We show, in fact, that (12) is a consequence of a far from innocent assumption that we shall refer to as *temporal additivity* which is to be contrasted with the more usual and different assumption of *causal additivity* made in clinical trials.

DOI: 10.1002/sim

4. MODELLING BASELINE AND OUTCOME

We now turn to consider the sort of model which will justify concluding that (12) applies and hence that SACS is unbiased whereas ANCOVA is biased. Suppose, as before, we index the two treatment groups by $i,\ i = A, C$, where A stands for active and C for control. Suppose we have n_A patients in the active treatment group and n_C in the control group and that a given patient may be indexed by $ij,\ i = A, C,\ j = 1, \ldots, n_i$. We let the expected values at baseline be

$$E[X_{ij}] = \mu + \theta_i \tag{13}$$

We now suppose that the expected value at outcome is

$$E[Y_{ij}] = \mu + \theta_i + \tau_i + \phi_i \tag{14}$$

Note that τ_A, τ_C are not separately identifiable and neither are ϕ_A, ϕ_B but that certain contrasts may be identifiable. For example, $\tau - \phi = (\tau_A - \tau_C) + (\phi_A - \phi_B)$ is identifiable from SACS and if it is assumed that $(\phi_A - \phi_B) = 0$ then τ is identifiable from this. Note that in terms of our discussion in Section 3, we have $B_X = (\theta_A - \theta_C)$, $B_Y = (\theta_A - \theta_C) + (\phi_A - \phi_C)$. Other contrasts may be identifiable given particular assumptions.

However, how should we interpret τ, θ and ϕ? The basic idea here is that τ reflects 'treatments' and θ the expected initial difference between groups. The ϕ parameters reflect 'secular trends'. However, without further examination, these designations are merely labels. How can we understand, for example, the meaning of τ? One possible way is in terms of the Rubin causal model [18]. We regard τ as the expected (unobservable) difference between the outcome if a patient is given treatment A rather than C. In practice a patient is only given either A or C and so one of the responses is counterfactual. Hence τ reflects differences between observables and counterfactuals. In theory such differences can be indexed by patient. Use of such counterfactuals has been criticized by Dawid [19].

For the moment, however, we avoid answering this difficult question and accept these particular parameter labels at face value without delving too closely into their meaning. Now, if we examine the papers of Liang and Zeger and Samuels in the light of the model for expected values given by (13) and (14) we can see that they effectively assume that $\phi = 0$. This is the *temporal additivity* assumption. It amounts to supposing that despite the fact that groups are different at baseline they would show the same evolution over time, that is to say $\phi_A = \phi_B$ in the absence of any effect of treatment. That being so, it does indeed follow that ANCOVA will be biased unless $\theta = 0$, that is to say unless the means at baseline are equal. This immediately raises the question, however, is it likely or even possible that circumstances can arise where

$$\theta \neq 0 \quad \text{and} \quad \phi = 0 \tag{15}$$

We discuss this point in Section 5 below, in which we show that for a trialist to assign patients to in such a way that

$$B_Y = \beta B_X$$

or

$$\begin{aligned} \theta + \phi &= \beta\theta \\ \phi &= (\beta - 1)\theta \end{aligned} \tag{16}$$

so that ANCOVA is unbiased is feasible and, indeed, that there is a wide class of design for which this is possible. However, we also show that such a trialist might find it difficult to assign patients in such a way that it is simultaneously true that (15) applies and a causal question of interest

DOI: 10.1002/sim

remains. This, in our view raises doubts as to whether SACS can ever have any utility in causal investigations.

5. ASSIGNMENT MECHANISMS

In this section we shall show that it is easy to design trials in which, given an assumption of causal additivity, the following conditions hold:

1. ANCOVA is unbiased;
2. SACS is biased;
3. A causal interpretation can be given.

However, we shall also show that it is rather difficult to design trials for which

1. SACS is unbiased;
2. ANCOVA is biased;
3. A causal interpretation can be given.

As regards the first set of conditions, we have already presented such a trial in Section 2. In fact any trial in which patients are assigned to treatment on the basis of their baseline value will yield valid estimates of 'the treatment effect' provided that ANCOVA is used, there is a linear relationhsip between baseline and outcome and the treatment effect is constant for all values of the baseline [20]. (That is to say there is no treatment by baseline interaction.) We refer to such an assumption as *causal additivity*. Of course, this assumption is not likely to be exactly true in practice but it is also not essential in order to make practical use of a *randomized* clinical trial. All that one needs is to estimate appropriately some average treatment effect and to assume that it brings some advantage over no estimate at all in predicting individual effects. See Senn [21] for a discussion of estimation in the presence of interaction and Senn [22] for a discussion of additivity. For a design in which patients have been allocated according to baseline in such a way that the two baseline means are not equal, the assumption of additivity is more important than for a randomized clinical trial [23] but note that this is a problem for the SACS estimator also.

In fact, such designs, in which subjects are assigned on the basis of baseline values are referred to as *cut-off designs* and have been extensively discussed in the educational literature [24] and even proposed in the medical literature [25]. Now, suppose that if all patients are given treatment i then baseline and outcomes are jointly Normally distributed so that

$$X, Y \sim \mathrm{N}(\mu_X, \mu_X + \tau_i, \sigma_X^2, \sigma_Y^2, \rho) \tag{17}$$

then, if we write

$$\mu_X = \mu, \quad E[X|X<k] = \mu + \theta_1, \quad E[X|X>k] = \mu + \theta_2 \tag{18}$$

the situation at baseline is as described by (13) and it is well known that we have

$$\begin{aligned} E[Y|X<k] &= \mu_Y + \tau_A + \rho\,\frac{\sigma_Y}{\sigma_X}\,\theta_1 = \mu_Y + \tau_A + \beta\theta_1 \\ E[Y|X\geqslant k] &= \mu_Y + \tau_C + \rho\,\frac{\sigma_Y}{\sigma_X}\,\theta_2 = \mu_Y + \tau_B + \beta\theta_2 \end{aligned} \tag{19}$$

DOI: 10.1002/sim

Hence the expected value of the contrast at outcome from (19) is

$$\tau_A - \tau_B + \beta(\theta_1 - \theta_2) = \tau + \beta\theta \tag{20}$$

Thus the amount by which (20) has to be corrected is not θ but $\beta\theta$. Hence ANCOVA is unbiased and SACS is biased.

Note that in estimating β an important assumption that makes ANCOVA unbiased is that the regression within groups is the same as that between, the latter being the potential bias and the former that by which the correction factor is estimated. Our opening example in Section 2 is a little unusual in this respect. The reduction to stratification on a binary covariate means that there is no extra information within groups.

Is it possible to construct examples in which SACS is unbiased but ANCOVA is not? Here are some apparently plausible schemes that do not work.

1. *Select patients according to their 'true' rather than observed values*: The idea behind this is that the origin of the regression effect that causes SACS to be biased is that the covariance of measurements is the variance of true values, say σ^2_{T}. The regression of outcome on observed baseline is $\sigma^2_{\mathrm{T}}/\sigma^2_{\mathrm{M}}$, where σ^2_{M}, $\sigma^2_{\mathrm{M}} > \sigma^2_{\mathrm{T}}$ is the variance of observed measurements. Hence the regression of outcome on baseline, which is the origin of the bias in SACS. However, if selection is on true values, then the relevant regression coefficient is $\sigma^2_{\mathrm{T}}/\sigma^2_{\mathrm{T}} = 1$ and SACS is unbiased. This argument does not work for two reasons. First, selection on the true value is impossible because it cannot be known. The second more serious objection is that true values at outcome and baseline will in any case not be perfectly correlated (after all, although given the same follow up all subjects age by the same amount during the study, not everybody's blood pressure rises with age at the same rate, so the correlation between blood pressure at baseline and outcome will not be perfect).
2. *Select patients on the basis of some continuous value other than the baseline measurement*: For example, we might be studying diastolic blood pressure but select patients on the basis of systolic blood pressure. Such an approach will indeed make ANCOVA biased unless the selecting covariate is itself included in the model or unless the partial correlation between selecting covariate and outcome is zero given the baseline. However, it will make SACS biased as well.
3. *Patients are selected on the average of their values at baseline and outcome*: This, would indeed, given additivity of the treatment effect, eliminate the regression effect. It was, for example, suggested by Oldham as being a way of studying baseline by treatment interaction for this very reason [26]. However, it is a logical impossibility as a selection mechanism, since it requires knowledge of future values at the time of selection.
4. *Select on a binary covariate*: The problem with this approach is that the binary covariate must not vary over time. If it does, then the influence of the binary covariate is not permanent. The proponent of SACS over ANCOVA is now on the horns of a dilemma. Either the binary covariate does not influence or mark any difference in measures between groups, in which case ANCOVA and SACS are valid or it does, in which case the difference between groups at outcome may be expected to be less than that at baseline and SACS is biased. The only way to get round this objection is to use a temporally unchanging covariate to allocate. For example, if in a comparison of two diets, females are assigned to one diet and males to another. This was, in fact, similar to the example Lord used to introduce his famous paradox [13]. The problem

DOI: 10.1002/sim

here however are first, that the assumption that the difference at outcome would be the same as the difference at baseline in the absence of treatment is rather dubious and second that it is debatable as to whether any causal interpretation of such an 'experiment' would be justified.

Of course, what the above show is merely that it is easy for trialists to design trials in which ANCOVA is valid and SACS is not but that the reverse is difficult. It may be that 'Nature' can design the latter trials easily. The point of this discussion however is not to propose arbitrary and inferior alternatives to randomized clinical trials but to try enquire what natural mechanisms might produce in the way of possible biases. In our opinion, careful consideration of this leads to the following conclusions. First, where an RCT has been run ANCOVA is a superior choice of analysis to SACS and will at least deal with accidental bias, whereas SACS will not. Second, that for other types of study, such as say epidemiological cohort studies it is, indeed plausible that ANCOVA will be biased but it is also plausible that SACS will be biased too. The practical implications of this will be considered in the next section.

6. DISCUSSION

In my view a number of lessons can be drawn from the above investigations.

The first is that one has to be rather critical and careful in specifying so-called repeated measures models. A natural tendency amongst modellers appears to be to model baselines and outcomes using an integrated framework. This is most extreme in the form of model made popular by Laird and Ware [27, 28] in which baselines are actually treated as being a *response* to treatment, which temporally and logically they cannot be [17, 29]. Usually the way in which estimation proceeds using such approaches means that a 'correct' estimate is recovered and no harm is done. Nevertheless, in my view there is a danger in assuming that it is a necessary feature of repeated measures designs that baselines should be modelled as if they were outcomes: it is unnatural from other points of view [17, 29]. Instead one should focus clearly on 'outcomes' as being the only values that can be influenced by treatment and examine critically any schemes that assume that these are linked in some rigid and deterministic view to 'baseline' values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way.

The second lesson is that treatment groups that differ at baseline do, indeed, present a challenge for any attempt at causal inference. The potential problems are well encapsulated by the famous paradox of Lord [13]. As noted in the introduction, a penetrating analyses of this paradox has been given by Holland and Rubin [14] and also by Wainer [15] more recently by Wainer and Brown [16]. Unfortunately, unlike these authors, many when they encounter the paradox assume that it implies that analysis of covariance is invalid but that analysis of change scores is not. The interpretation offered here is more radical: ANCOVA will often be biased under such circumstances but SACS almost certainly so.

The third lesson is more general. You cannot establish necessary conditions for an estimator to be valid by nominating a model and seeing what the model implies unless the model is universally agreed to be impeccable. On the contrary it is appropriate to start with the estimator and see what assumptions are implied by valid conclusions.

ACKNOWLEDGEMENTS

I thank Emmanuel Lesaffre and Mike Kenward for helpful discussions and the referees for helpful comments.

DOI: 10.1002/sim

REFERENCES

1. Senn SJ. The use of baselines in clinical trials of bronchodilators. *Statistics in Medicine* 1989; **8**:1339–1350.
2. Senn SJ. Covariate imbalance and random allocation in clinical trials [see comments]. *Statistics in Medicine* 1989; **8**:467–475.
3. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design [see comments]. *Statistics in Medicine* 1992; **11**:1685–1704.
4. Frison LJ, Pocock SJ. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine* 1997; **16**:2855–2872.
5. Chambless LE, Roeback JR. Methods for assessing difference between groups in change when initial measurements is subject to intra-individual variation. *Statistics in Medicine* 1993; **12**:1213–1237.
6. Senn SJ. Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation [letter; comment] [see comments]. *Statistics in Medicine* 1994; **13**:2280–2285.
7. Senn SJ. In defence of analysis of covariance: a reply to Chambless and Roeback [letter; comment]. *Statistics in Medicine* 1995; **14**:2283–2285.
8. Senn SJ. Covariance analysis in generalized linear measurement error models [letter; comment]. *Statistics in Medicine* 1990; **9**:583–586.
9. Senn SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design [letter; comment]. *Statistics in Medicine* 1994; **13**:197–198.
10. Liang KY, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya—The Indian Journal of Statistics Series B* 2000; **62**:134–148.
11. Samuels ML. Use of analysis of covariance in clinical trials: a clarification. *Controlled Clinical Trials* 1986; **7**:325–329.
12. Laird N. Further comparative analyses of pre-test post-test research designs. *The American Statistician* 1983; **37**:329–330.
13. Lord FM. A paradox in the interpretation of group comparisons. *Psychological Bulletin* 1967; **66**:304–305.
14. Holland PW, Rubin DB. On Lord's Paradox. In *Principals of Modern Psychological Measurement*, Wainer H, Messick S (eds). Lawrence Erlbaum Associates: Hillsdale, NJ, 1983.
15. Wainer H. Adjusting for differential base rates—Lords Paradox again. *Psychological Bulletin* 1991; **109**:147–151.
16. Wainer H, Brown LM. Two statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data. *American Statistician* 2004; **58**:117–123.
17. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Chichester, 1997.
18. Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
19. Dawid AP. Causal inference without counterfactuals. *Journal of the American Statistical Association* 2000; **95**:407–424.
20. Rubin DB. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 1977; **2**: 1–26.
21. Senn SJ. The many modes of meta. *Drug Information Journal* 2000; **34**:535–549.
22. Senn SJ. Added values: controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* 2004; **23**:3729–3753.
23. Senn SJ. A personal view of some controversies in allocating treatment to patients in clinical trials [see comments]. *Statistics in Medicine* 1995; **14**:2661–2674.
24. Campbell DT, Stanley JC. *Experimental and Quasi-experimental Designs for Research*. Rand McNally: Chicago, 1966.
25. Trochim WM, Cappelleri JC. Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials* 1992; **13**:190–212.
26. Oldham P. A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases* 1962; **15**:969–977.
27. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
28. Laird NM, Wang F. Estimating rates of change in randomized clinical trials. *Controlled Clinical Trials* 1990; **11**:405–419.
29. Senn SJ, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine* 2000; **19**:861–877.

DOI: 10.1002/sim