Community Dent Oral Epidemiol 2007; 35: 412–420

Community Dentistry and Oral Epidemiology

Methods

Statistical issues on the analysis of change in follow-up studies in dental research

Andrew Blance[1,2], Yu-Kang Tu[1,2], Vibeke Baelum[3] and Mark S. Gilthorpe[2]

[1]Leeds Dental Institute, University of Leeds, Clarendon Way, Leeds, UK, [2]Biostatistics Unit, Centre for Epidemiology & Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds, UK, [3]Department of Community Oral Health, Faculty of Health Sciences, University of Aarhus, Aarhus C, Denmark

Blance A, Tu Y-K, Baelum V, Gilthorpe MS. Statistical issues on the analysis of change in follow-up studies in dental research. Community Dent Oral Epidemiol 2007; 35: 412–420. © 2007 The Authors. Journal compilation © 2007 Blackwell Munksgaard

Abstract – *Objective:* To provide an overview to the problems in study design and associated analyses of follow-up studies in dental research, particularly addressing three issues: treatment-baselineinteractions; statistical power; and nonrandomization. *Background:* Our previous work has shown that many studies purport an interacion between *change* (from baseline) and *baseline* values, which is often based on inappropriate statistical analyses. *A priori* power calculations are essential for randomized controlled trials (RCTs), but in the pre-test/post-test RCT design it is not well known to dental researchers that the choice of statistical method affects power, and that power is affected by treatment-baseline interactions. A common (good) practice in the analysis of RCT data is to adjust for baseline outcome values using ANCOVA, thereby increasing statistical power. However, an important requirement for ANCOVA is there to be no interaction between the groups and baseline outcome (i.e. effective randomization); the patient-selection process should not cause differences in mean baseline values across groups. This assumption is often violated for nonrandomized (observational) studies and the use of ANCOVA is thus problematic, potentially giving biased estimates, invoking Lord's paradox and leading to difficulties in the interpretation of results. *Methods:* Baseline interaction issues can be overcome by use of statistical methods; not widely practiced in dental research: Oldham's method and multilevel modelling; the latter is preferred for its greater flexibility to deal with more than one follow-up occasion as well as additional covariates To illustrate these three key issues, hypothetical examples are considered from the fields of periodontology, orthodontics, and oral implantology. *Conclusion:* Caution needs to be exercised when considering the design and analysis of follow-up studies. ANCOVA is generally inappropriate for nonrandomized studies and causal inferences from observational data should be avoided.

Key words: analysis of change; analysis of covariance (ANCOVA); Lord's paradox; mathematical coupling; statistical epidemiology

Prof. Mark S Gilthorpe, Biostatistics Unit, Centre for Epidemiology & Biostatistics, LIGHT, University of Leeds, 30-32 Hyde Terrace, Leeds LS2 9LN, UK
Tel: +(44) 113 343 1913
Fax: +(44) 113 343 4877
e-mail: m.s.gilthorpe@leeds.ac.uk

Submitted 6 November 2006;
accepted 24 April 2007

Introduction

The analysis of change in outcome variables measured on two or more occasions is one of the most commonly used designs in follow-up studies in oral health research. Results from the test group are compared with those from the control group, to show whether or not changes in the outcome differ statistically significantly. There are several discussions in the oral health literature on appropriate statistical methods for follow-up studies; the concepts and rationale behind the general follow-up study design therefore seem quite straightforward. However, despite warnings about the misuse of statistical methods in the analysis of change (1, 2), certain inappropriate practices continue.

Oral health researchers frequently overlook three key issues in the analyses of change: (i) treatment–baseline interactions, referred to as 'baseline effects'; (ii) statistical power and (iii) non-randomization. While the power of follow-up studies is extensively addressed within the medical and

doi: 10.1111/j.1600-0528.2007.00407.x

statistical literature (3, 4), the impact of baseline effects, particularly on statistical power, is not widely appreciated. Furthermore, insufficient consideration is given to the choice of statistical methods and their consequences in non-randomized studies, particularly as non-randomization reverses otherwise standard advice to use ANCOVA to generally maximize statistical power within randomized controlled trials (RCTs).

The aim of this article is to provide a non-technical introduction to the current problems in study design and associated analyses of follow-up studies in oral health research, particularly addressing the issues of: baseline effects, power and non-randomization.

Baseline effects

Many studies in the dental literature show an association between *baseline* outcome status and *change* from baseline, i.e. a treatment–baseline interaction or *baseline effect*. For instance, in periodontal follow-up studies, probing pocket depth (PPD) reductions and clinical attachment level (CAL) gains have often been found to be positively associated with baseline measurements of PPD and CAL (5, 6). Similarly, the effect of orthodontic treatment of malocclusions, assessed as changes in the peer assessment rating (PAR) score, has been found to be positively associated with pre-treatment PAR scores (7–12). This is not unique to dentistry, and examples are found in studies of hypertension treatments, showing that patients with higher-than-average blood pressure might experience greater blood pressure reduction following a pharmacological intervention than those with baseline blood pressures lower than the study average (13).

The problem is that use of correlation or regression to test the association between *change* in an outcome and its *baseline* value suffers a serious statistical artefact: mathematical coupling (MC) (14, 15). MC occurs where there exists a formulaic relationship between two variables, i.e. one can be expressed as a function of the other. MC distorts the perceived relationship between variables, as the usual statistical testing of the null hypothesis – i.e. that the correlation coefficient or regression slope is zero – becomes inappropriate. For instance, suppose, that pre-treatment PPD is x_1, post-treatment PPD x_2, and therefore PPD reduction following treatment is $x_1 - x_2$. To correlate (or regress) $x_1 - x_2$ with x_1 may invalidate the usual null hypothesis because x_1 appears in both variables. Any association between $x_1 - x_2$ and x_1 (i.e. a non-zero statistical correlation between $x_1 - x_2$ and x_1) may exist, in part, because of MC (as $x_1 - x_2$ and x_1 are formulaically related). For instance, if x_1 and x_2 were two series of random numbers with the same mean and standard deviation, the expected correlation between x_1 and x_2 is close to zero. However, it can be shown that the correlation between $x_1 - x_2$ and x_1 in such circumstances will be close to $1/\sqrt{2} \approx 0.71$ (16). This value can be highly significant when tested against the (incorrect) null hypothesis of zero, even with a small sample size (14). Researchers may thereby be misled to infer an underlying 'causal' relationship between $x_1 - x_2$ and x_1, where none exists.

Problematic uses of correlation and/or regression in analysing the association between treatment effects and baseline values have been noted for a long time now (17–20). One needs to know the correct null hypothesis, and a method has been proposed to obtain an estimate of this (21). However, this approach does not provide a gauge of the extent of association, as provided by a correlation coefficient. Alternatively, over 40 years ago, Oldham (17) suggested that one solution was to test the correlation of $x_1 - x_2$ with the average $(x_1 + x_2)/2$. The reason is that to know whether or not a baseline effect exists, a statistically correct approach is to test for differences in the variances of the two measurements, rather than to test the correlation coefficient between *change* and *baseline*. In the latter (erroneous) approach, MC adds to and exacerbates the statistical artefact known as regression to the mean (RTM) (22, 23), as the variables and also their measurement errors are formulaically related. A more technical explanation of this is outlined in more detail elsewhere (15).

It is important to note that Oldham's method does not remove MC, rather it uses the fact that if there is a baseline effect, the follow-up measurements will vary differently from the baseline measurements, because of the fact that the baseline effect will decrease the value of the observations. For example, consider periodontal treatment where a baseline effect means that initially deeper pocket depths (PPD) reduce (improve) more than initially shallower pockets. In statistical terms, this means that the follow-up measurements have a smaller standard deviation than the baseline measurements. An illustration adopting vector geometry is presented here, while the statistical theory is

provided in brief in Appendix 1. For readers not wishing to consider vector geometry, the following section can be omitted without loss of continuity.

Using vector geometry (24), pre-treatment (x_1) and post-treatment (x_2) PPD values can be represented as vectors with lengths equal to their standard deviations (SDs), positioned such that the cosine of the angle between them is their bivariate correlation (Fig. 1). Under the null hypothesis (H_0) of no baseline effect, the SD of pre- and post-treatment values should be equal (15, 17). In vector-geometrical terms this means that the two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are perpendicular, and their lengths the same. (Note: we use bold to distinguish between vector representation of the variables and their usual variable representation.) The correlation between change ($x_1 - x_2$) and baseline (x_1) is now equivalent to the cosine of the angle ψ between the vectors $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_1$. This is typically not zero, but depends upon the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e. the correlation between pre- and post-treatment PPD. When this is near zero, i.e. the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are perpendicular, the angle between $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_1$ is 45°, and its cosine is $1/\sqrt{2} \approx 0.71$. Thus, under H_0, the correlation between change and baseline is generally far from the standard assumption of a value of zero, and this distortion to the null hypothesis is a consequence of MC.

Vector geometry also illustrates the rationale behind Oldham's method. A relationship between change and baseline requires that the variances of the baseline and follow-up measurements are unequal, i.e. one standard deviation is smaller than the other. In vector geometrical terms, this would mean unequal lengths of $\mathbf{x}_1$ and $\mathbf{x}_2$. However, under H_0 (of no underlying relation), the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are of the same length, and the vectors $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_1 + \mathbf{x}_2$ are therefore always perpendicular, irrespective of the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ (Fig. 1). Thus, the correlation between $x_1 - x_2$ (change) and $(x_1 + x_2)/2$ (mean) is always zero, under H_0. Therefore, although MC remains, the use of Oldham's method under H_0 provides a special instance where its adverse effect (i.e. distortion to the null hypothesis) is annulled.

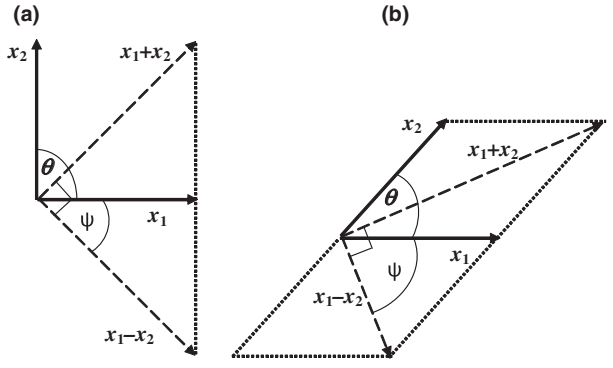

Fig. 1. Variables x_1 (baseline PPD) and x_2 (follow-up PPD) represented as vectors with lengths equal to their standard deviation (SD); cosine θ is the correlation between x_1 and x_2; under H_0 (the SDs of x_1 and x_2 are equal) the vectors $x_1 - x_2$ and $x_1 + x_2$ are always perpendicular, irrespective of the correlation between x_1 and x_2: (a) the correlation between x_1 and x_2 is zero, hence $\theta = 90°$ and $\psi = 45°$; (b) the correlation between x_1 and x_2 is positive, hence $\theta < 90°$ and $\psi > 45°$. MC is still present but use of Oldham's method annuls the effects.

Modelling baseline effects

Simple statistical methods, such as Oldham's correlation (17), have been recommended to overcome the problem of testing the interaction between treatment effects and baseline values. However, these methods have limited applications. For instance, Oldham's method assumes that measurement errors are constant across occasions, and cannot take into consideration other explanatory variables, such as treatment group variables. An alternative approach would be to use multilevel modelling (MLM) (25, 26), which is more flexible in dealing with repeated measurement data, avoids the problems of MC, permits multiple follow-up time-points, and permits the inclusion of additional covariates (27–31). MLM is more complex than Oldham's method, so we only outline the basic principles here for the pre-/post-test study design; more technical details and discussion are given in Appendix 2 and elsewhere (32).

The MLM required to analyse *change* in relation to *baseline*, while completely avoiding MC, is where one specifies baseline and follow-up values as repeated outcomes (at level 1) clustered within individuals (at level 2). Within this model, measurement occasion is a covariate, where its coefficient exhibits random variation about its mean (26). This is known as a *random slope* model because the estimated slope (randomly) varies across individuals (level 2) (33). The occasion covariate is centred about 0 to aid model-fitting procedures (32), and its interval, though arbitrary, is set to 1 so that interpretation of its regression coefficient becomes the *mean change* between occasions. The random structure of the model comprises subject-level random intercept, subject-level random slope and a covariance between them, which is used to derive

the correlation between baseline (intercept) and change (slope) (34), free from the distortion due to MC. This strategy can be extended to accommodate observations of multiple sites (e.g. treatment of different lesions) within the same individual, by including an extra level for site. Thus, baseline and follow-up values (level 1) are clustered within sites (level 2), which in turn are clustered within individuals (level 3). Furthermore, other factors such as treatment group may be incorporated in the MLM as additional covariates. More complex variations can also be developed to consider multiple follow-up measures, though this is beyond the scope of this article to outline these. As a simple example, consider orthodontic PAR scores to evaluate the effect of orthodontic treatment of malocclusions: baseline and follow-up PAR scores form level 1 observations nested within subjects at level 2. PAR scores from both occasions are regressed on the occasion covariate and its coefficient is allowed to exhibit random variation about an overall mean value. MC is not present in this model as the dependent variable has no formulaic relationship with the independent variable.

Statistical power

When conducting a randomized controlled trial (RCT), *a priori* power calculations are necessary to determine the required sample size. This is often overlooked or under-reported in the oral health literature (35). In the repeated measurement study design, typically adopted by RCTs, it is not well known among oral health researchers that the analytical method of choice affects statistical power. Moreover, the power of most statistical methods to analyse repeated measurement designs are affected by baseline effects.

In a separate study (36), only summarized here, computer simulations were performed to compare the power of four univariate statistical methods and two multivariate statistical methods for the analysis of change in a hypothetical RCT involving two measurements, one at baseline and the other at follow-up. The univariate methods considered were: (a) testing post-treatment scores only using the two-sample *t*-test; (b) testing change scores using the two-sample *t*-test; (c) testing percentage change scores using the two-sample *t*-test; and (d) analysis of covariance (ANCOVA). The two multivariate methods considered were: (e) multivariate analysis of variance (MANOVA); and (f) multilevel modelling (MLM). All simulations were undertaken initially assuming that treatment effects were not related to baseline values (i.e. there was no baseline effect) and repeated assuming that treatment effect would increase for higher baseline values (i.e. there was a baseline effect). In general, ANCOVA proved to be the most powerful method and always had greater power than the other commonly used methods such as change scores and percentage change scores. The two multivariate methods did not achieve greater power than ANCOVA (37).

Many statisticians claim that ANCOVA *always* achieves the greatest power unless the correlation between the pre- and post-treatment measures is zero, at which point ANCOVA achieves the same power as using post-treatment values only (3). However, this is true only when the sample size is 'reasonably' large. In our simulations (36), it was noted that ANCOVA might achieve less power than testing post-treatment values only, where the correlation between the pre- and post-intervention measurements was low (≤0.3), corresponding to varying treatment effect across individuals, and the sample size was small (≤20). The reason for this is that ANCOVA uses baseline values as a covariate and thus loses one degree of freedom more than the other methods; for small sample sizes, one degree of freedom can have a substantial effect if the correlation between pre- and post-treatment values is also small. Given that the average sample size of RCTs in oral health research is quite small (36), this finding might be important. Otherwise, in general, ANCOVA is the preferred method of analysis for reasonably sized RCTs, as this yields optimal statistical power.

ANCOVA and Lord's paradox

Although ANCOVA is recommended for RCT data (3, 4, 37), and is typically described as useful because it 'adjusts for baseline differences', the implicit assumption underlying ANCOVA is often overlooked or misunderstood. Consequently, many researchers have developed the naïve view that ANCOVA adjusts for baseline differences *between* groups, when the reality is that it adjusts only for baseline differences *within* groups. ANCOVA achieves this adjustment within treatment groups by using baseline values as a covariate, and it is this

that increases statistical power. If there is an interaction between baseline values and treatment groups, i.e. the patient selection process causes differences in the baseline values between treatment groups, the assumptions for ANCOVA may be violated and subsequent conclusions drawn could be erroneous.

For RCTs, no substantial differences in the mean baseline values across groups should exist, because (appropriate) randomization ensures that the distributions of baseline variables are very similar. In reality, small differences might be found, though these are assumed to be caused by chance alone and will not bias the ANCOVA estimates. By implication, within observational studies, i.e. where randomization is not performed, or randomization is not conducted appropriately, using ANCOVA to adjust for baseline differences could mislead by introducing bias into the ANCOVA estimates, giving rise to Lord's paradox (38) and yielding difficulties in the interpretation of results.

Lord's paradox occurs where baseline differences cannot be attributed to chance alone. Lord's paradox dictates that in instances where real baseline differences exist, it is erroneous to attempt to adjust for baseline differences, because ANCOVA has the potential to yield biased estimates of treatment differences (see Fig. 2). The original example described by Lord (38) is where one examines differences between males and females in the changes in body mass. Suppose, for instance, we wish to know if a special diet has a differential impact by sex on weight loss. Males and females will have different mean body mass at baseline and this cannot be attributed to chance, as sex cannot be assigned randomly. Controlling for baseline body mass in this instance is questionable and will invoke Lord's paradox.

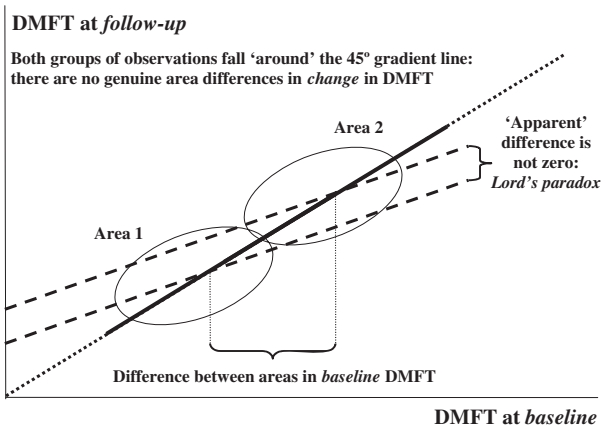

Fig. 2. Plot of baseline DMFT (x_1) versus follow-up DMFT (x_2) for children in a follow-up study of 5 years, following an implicit 'intervention' of water fluoridation in one area.

To visually explain this phenomenon with respect to follow-up studies, consider an investigation into the effect of water fluoridation on dental caries (DMFT) increments. Researchers might use data retrospectively or prospectively, collected from one geographical area with water fluoridation and another without fluoridation. Suppose repeated oral examinations are performed on children in both areas at an interval of 5 years and there are substantial differences in baseline caries rates. Even if the two areas had been randomly selected from fluoridated and non-fluoridated areas, there would remain the possibility that baseline differences occur due to the lack of 'appropriate' randomization. Here we imply that appropriate randomization warrants random allocation of fluoridation to previously non-fluoridated areas – which is not the same as randomly selecting fluoridated and non-fluoridated areas. Moreover, even with appropriate random allocation of fluoridation to areas, the study sample size would be only two! The problem is whether or not the DMFT-increment over the 5-year period can be compared between the two areas; and if there is a significant difference between areas, can this be attributed to water fluoridation?

Many researchers might seek some form of statistical adjustment for differences in baseline DMFT. However, this would be inappropriate. Under the null hypothesis (H_0) of the same 5-year change in DMFT among all children (i.e. irrespective of area), without any biological variation and/or measurement error, the follow-up DMFT (x_1) plotted against baseline DMFT (x_2) would yield a straight line (Fig. 2; the 45° dotted line). However, because of biological variation and/or measurement error, the reality is that the data form a 'cloud' of points around the 45° incline. Furthermore, as baseline DMFT differs between areas, there are two such 'clouds', one for each area (Fig. 2: the data points form ellipsoids). Because of RTM, the slope of the fitted line for follow-up DMFT regressed on baseline DMFT is not coincident with the 45° incline. Thus, the ANCOVA estimate of the difference between fluoridated and non-fluoridated areas is not zero, as required under H_0. This artefactual effect of area on the changes in DMFT (which some could erroneously interpret as being due to fluoridation) is due to RTM, thereby yielding Lord's paradox.

Contradictory findings from different statistical methods

To illustrate contradictory results that might be obtained using different statistical methods, where Lord's paradox occurs for non-randomized (observational) studies, consider a hypothetical example from oral implantology. Dental surgeons wish to know whether or not using a membrane barrier in conjunction with bone grafts gives a better resolution of dehiscence defects in immediate oral implantation compared to using bone grafts alone. To address this research question, the clinical team search their patients' records and find 30 cases with a combination treatment (membrane plus bone grafts) and 30 cases with bone grafts alone. Crucially, the decision to use a membrane or not was based on each clinicians' clinical judgement during the surgery and was not therefore randomized. The study data were submitted to two statisticians.

The first statistician was informed and compared the dehiscence fill (i.e. the change) between the two treatment groups using the two-sample *t*-test. The mean difference was 1.4 mm (95% CI 0.9–1.9) and the conclusion was that combination therapy achieved better outcomes than bone grafts alone (Table 1). The second statistician spotted that there was a difference in the baseline defect depth between treatment groups. To adjust for this imbalance in baseline defect depth, ANCOVA was used, taking post-treatment defect depth as the outcome and baseline defect depth as a covariate. The results showed that after adjustment for the imbalance in baseline defect depth the difference in treatment effects was no longer statistically significant (Table 1).

The problem of using ANCOVA in this simulated example is that the cause of imbalance in baseline defect depth between the two treatment groups is unclear. The allocation of patients to the treatment groups was not random, but based on clinicians' judgement, where the greater baseline defect depth in the combination therapy group could be due to clinicians believing that greater defects need membranes to create space for regeneration or prevent the loss of the bone graft. The imbalance in baseline defect depth might therefore be just one of many differences in the defect characteristics between the two groups, because of the selection process of the patients. This hypothetical study reveals how the use of ANCOVA for non-randomized (observational) studies can give rise to difficulties in the interpretation. This also reminds us that making causal inferences from observational (non-randomized) data should be very cautious (if not avoided).

Table 1. Summary of hypothetical data in a study to test the difference in the treatment efficacy (mm) between combination therapy (group 1) and single therapy (group 2)

	Group 1 (membrane + bone graft)		Group 2 (bone graft only)	
	Mean	SD	Mean	SD
Baseline defect depth	3.90	0.89	2.37	1.07
Follow-up depth	2.67	0.92	1.27	1.17
Change in depth	1.23	1.17	1.10	0.80

Concluding remarks

This article highlights several common problems in the analyses of data from follow-up studies in oral health research. Although these problems are well known within the statistical sciences, with most of the issues surrounding power well documented in the general medical literature, the aspects of baseline effects affecting power are relatively unknown. Furthermore, the dental research community frequently overlooks the problems associated with ANCOVA for non-randomized study designs. Consequently, some of the evidence purported in the oral health science literature needs to be (re-) evaluated with caution. Oral health researchers need to be aware of these potential problems in study design and associated data analyses to avoid generating misleading evidence in the future.

Appendix 1

Suppose that within a follow-up study, pre-treatment PPD is x_1, post-treatment PPD x_2, it can be shown that the Pearson correlation between the *change* ($x_1 - x_2$) and the pre-treatment *baseline* value (x_1) is (17):

$$r_{x1-x2,x1} = \frac{\sigma_1 - r_{12}\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2}}; \qquad (A1)$$

where σ_1^2 is the observed variance (standard deviation squared) of x_1, $\sigma_2^2$the observed variance of x_2, and r_{12} the correlation between x_1 and x_2. Because of measurement errors or heterogeneous response

to the treatment, giving rise to RTM, the correlation between baseline and post-treatment values (r_{12}) will be smaller than 1, and, therefore, the correlation between *change* and *baseline* tends to be greater than 0, unless $\sigma_2^2$is much greater than σ_1^2. It is directly the consequence of the formulaic relationship between $x_1 - x_2$ and x_1 (i.e. MC) that the numerator in A1 depends upon r_{12}, and when this is not unity (i.e. when RTM operates), the usual null hypothesis of zero correlation is effectively 'distorted' (away from zero). Consequently, to correctly test the association between *change* and *baseline*, the impact of RTM needs to be estimated and then explicitly accommodated, and, unfortunately, this is not always achievable.

The Pearson correlation coefficient for Oldham's method is given by (17):

$$r_{x1-x2,(x1+x2)/2} = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4r_{12}\sigma_1^2\sigma_2^2}}. \qquad \text{(A2)}$$

Clearly, the correlation between $(x_1 - x_2)$ and $(x_1 + x_2)/2$ will be zero if the variances of x_1 and x_2 are equal, and positive if and only if $\sigma_1^2$is greater than σ_2^2.The impact of MC has been annulled, even though there remains a formulaic relationship between $(x_1 - x_2)$ and $(x_1 + x_2)/2$ because, in this special instance, the numerator of equation (A2) no longer contains r_{12}, and is therefore unaffected by RTM (when r_{12} is not unity).

Appendix 2

In order to illustrate the MLM approach in determining a baseline effect (i.e. interaction between *change* following treatment and *baseline*), while completely avoiding MC, consider the use of orthodontic PAR scores to evaluate the malocclusion of patients pre- and post-treatment. Baseline and follow-up PAR scores form level 1 observations ($i = 1, 2$) nested within subjects at level 2 ($j = 1, \ldots, N$), where N is the number of study subjects. PAR scores from both occasions are then regressed on the occasion covariate, say T, which is centred about zero [to avoid inducing the bias: see Blance et al. (32) for details] and adopts values such that it spans an interval of one ($T = \pm 1/2$). The coefficient for T exhibits random variation about its mean, yielding a multilevel regression model of the following form:

$$\mathrm{PAR}_{ij} = \mathrm{B}_{0ij} + \mathrm{B}_{1j}\mathrm{T}; \quad \mathrm{B}_{0ij} = \mathrm{B}_0 + \mathrm{u}_{0j} + \mathrm{e}_{0ij}; \quad \mathrm{B}_{1j} = \mathrm{B}_1 + \mathrm{u}_{1j},$$

where B_0 is the mean intercept of the sample values at a time point midway between *baseline* (pre-intervention) and follow-up (post-intervention); B_1 is the slope of the *change* in PAR score between measurement occasions; u_{0j} is the residual variation for individual j about the mean intercept due to population biological variation (heterogeneity between individuals of a population); e_{0ij} is the residual variation for individual j about the mean outcome on measurement occasion i, due to instantaneous biological variation (variation within an individual) and/or measurement error (which may differ between occasions though it is assumed at least for this illustration to be independent across occasions); u_{1j} is the responsive biological variation between subjects, i.e. the variation of the regression slope; and all variation is assumed to be normally distributed with zero mean.

Allowing for instantaneous biological variation and/or measurement error to differ across measurement occasions, there are five random parameters to be estimated, yet only three degrees of freedom: one for each occasion and one between occasions (change). It is therefore necessary to reduce the number of random parameters by making various model assumptions. The final model is contingent on these assumptions. For instance, if we were to acknowledge that we are unable to distinguish between population biological variation and instantaneous biological variation and/or measurement error, and we further assume the latter to have constant variance across occasions, we may estimate either the subject-level random intercept or the occasion-level random intercept, though not both. It does not affect our interpretation of the model whichever we choose (constraining the other to be zero), as the chosen estimate represents the combined effects of population and instantaneous biological variation and measurement error across the study period.

While the MLM strategy removes MC, it does not remove the impact of measurement error, i.e. the remaining effects of RTM. However, if an estimate of the error variance were obtained (or estimated), adjustment can then be made for the effects of measurement error. Although the working details of this are beyond the scope of this article, it can be shown that, providing the measurement error variance is constant across

occasions and it is independent of both population variation (random intercept) and responsive variation (random slope), the error-free correlation between *baseline* and *change*, P_{u01}, is related to the observed correlation, ρ_{u01}, according to formula $P_{u01} = \rho_{u01}\sqrt{1+R}$,where R is the ratio of the error variance to that of the population variance (the latter is estimated from the sample). Under these restricted assumptions, the observed correlation between *change* and *baseline* is always biased towards zero due to measurement error, as $\sqrt{1+R} \geq 1$. More complex MLMs can address non-constant biological variation and/or measurement error across multiple follow-up occasions. The formulation of models where variation is dependent on unobserved outcomes is ongoing.

References

1. Macfarlane TV, Worthington HV. Some aspects of data analysis in dentistry. Community Dent Health 1999;16:216–9.
2. Tu YK, Maddick IH, Griffiths GS, Gilthorpe MS. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. J Dent 2004;32:133–42.
3. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
4. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. Br Med J 2001;323:1123–4.
5. Lindhe J, Karring T, Lang NP, editors. Clinical periodontology and implant dentistry, 4th edn. London: Munksgaard/Blackwell; 2003.
6. Cortellini P, Tonetti MS. Focus on intrabony defects: guided tissue regeneration. Periodontol 2000 2000;22:104–32.
7. Shaw WC, Richmond S, O'Brien KD, Brook P, Stephens CD. Quality control in orthodontics: indices of treatment need and treatment standards. Br Dent J 1991;170:107–12.
8. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Jones R, Stephens CD et al. The development of the PAR index (Peer Assessment Rating): reliability and validity. Eur J Orthod 1992;14:125–39.
9. Richmond S, Shaw WC, Roberts CT, Andrews M. The PAR index (Peer Assessment Rating): methods to determine outcome of orthodontic treatment in terms of improvement and standards. Eur J Orthod 1992;14:180–7.
10. DeGuzman L, Bahiraei D, Vig KW, Vig PS, Weyant RJ, O'Brien K. The validation of the Peer Assessment Rating index for malocclusion severity and treatment difficulty. Am J Orthod Dentofacial Orthop 1995;107: 172–6.
11. Kerr WJ, Buchanan IB, McColl JH. Use of the PAR index in assessing the effectiveness of removable orthodontic appliances. Br J Orthod 1993;20:351–7.
12. John W, Kerr S, Buchanan IB, McNair FI, McColl JH. Factors influencing the outcome and duration of removable appliance treatment. Eur J Orthod 1994;16:181–6.
13. Gill JS, Zezulka AV, Beevers DG, Davies P. Relation between initial blood pressure and its fall with treatment. Lancet 1985;325:567–9.
14. Tu YK, Gilthorpe MS, Griffiths GS. Is reduction of pocket probing depth correlated with the baseline value or is it 'mathematical coupling'?. J Dent Res 2002;81:722–6.
15. Tu Y-K, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. Stat Med 2007;26:443–57.
16. Andersen B. Methodological errors in medical research. London: Blackwell; 1990.
17. Oldham PD. A note on the analysis of repeated measurements of the same subjects. J Chronic Dis 1962;15:969–77.
18. Blomqvist N. On the relation between change and initial value. J Am Stat Assoc 1977;72:746–9.
19. Altman DG. Statistics in medical journals. Stat Med 1982;1:59–71.
20. Altman DG. Statistics in medical journals: developments in the 1980s. Stat Med 1991;10:1897–1913.
21. Tu YK, Baelum V, Gilthorpe MS. The relationship between baseline value and its change: problems in categorization and the proposal of a new method. Eur J Oral Sci 2005;113:279–88.
22. Kirkwood BR, Sterne JA. Essential medical statistics. London: Blackwell; 2003.
23. Egelberg J. The impact of regression towards the mean on probing changes in studies on the effect of periodontal therapy. J Clin Periodontol 1989;16: 120–3.
24. Wickens TD. The geometry of multivariate statistics. Hillsdale: Lawrence Erlbaum Associates; 1995.
25. Gilthorpe MS, Griffiths GS, Maddick IH, Zamzuri AT. The application of multilevel modelling to periodontal research data. Community Dent Health 2000;17:227–35.
26. Gilthorpe MS, Griffiths GS, Maddick IH, Zamzuri AT. An application of multilevel modelling to longitudinal periodontal research data. Community Dent Health 2001;18:79–86.
27. Gunnell D, Berney L, Holland P, Maynard M, Blane D, Davey Smith G et al. Does the misreporting of adult body size depend upon an individual's height and weight? Methodological debate Int J Epidemiol 2004;33:1398–9.
28. Gilthorpe MS, Tu YK. Mathematical coupling: a multilevel approach. Int J Epidemiol 2004;33:1399–1400.
29. White IR. Assessing correlation between reporting errors and true values: untestable assumptions are unavoidable. Int J Epidemiol 2004;33:1400–1.
30. Rasbash J, Goldstein H. Mathematical coupling: a simpler approach. Int J Epidemiol 2004;33:1401–2.
31. Gilthorpe MS, Tu YK, Gunnell D. A coda: oversimplification, implicit assumptions, and measurement error. Int J Epidemiol 2004;33:1402–3.

32. Blance A, Tu YK, Gilthorpe MS. A multilevel modelling solution to mathematical coupling. Stat Methods Med Res 2005;14:553–65.
33. Gilthorpe MS, Zamzuri AT, Griffiths GS, Maddick IH, Eaton KA, Johnson NW. Unification of the 'burst' and 'linear' theories of periodontal disease progression: a multilevel manifestation of the same phenomenon. J Dent Res 2003;82:200–5.
34. Bryk AS, Raudenbush AW. Hierarchical linear models: applications and data analysis methods. London: Sage; 1992.
35. Tu YK, Maddick I, Kellett M, Clerehugh V, Gilthorpe MS. Evaluating the quality of active-control trials in periodontal research. J Clin Periodontol 2006;33: 151–6.
36. Tu YK, Blance A, Clerehugh V, Gilthorpe MS. Statistical power for analyses of changes in randomized controlled trials. J Dent Res 2005;84:283–7.
37. Bonate P. Analysis of pretest–posttest designs. Boca-Raton, FL: Chapman & Hall/CRC; 2000.
38. Lord FM. A paradox in the interpretation of group comparisons. Psychol Bull 1967;68:304–5.