ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 920–925

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies

Gerard J.P. Van Breukelen*

Department of Methodology & Statistics, Research Institute Caphri, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Accepted 13 July 2005

Abstract

Background and Objective: For inferring a treatment effect from the difference between a treated and untreated group on a quantitative outcome measured before and after treatment, current methods are analysis of covariance (ANCOVA) of the outcome with the baseline as covariate, and analysis of variance (ANOVA) of change from baseline. This article compares both methods on power and bias, for randomized and nonrandomized studies.

Methods: The methods are compared by writing both as a regression model and as a repeated measures model, and are applied to a nonrandomized study of preventing depression.

Results: In randomized studies both methods are unbiased, but ANCOVA has more power. If treatment assignment is based on the baseline, only ANCOVA is unbiased. In nonrandomized studies with preexisting groups differing at baseline, the two methods cannot both be unbiased, and may contradict each other. In the study of depression, ANCOVA suggests absence, but ANOVA of change suggests presence, of a treatment effect. The methods differ because ANCOVA assumes absence of a baseline difference.

Conclusion: In randomized studies and studies with treatment assignment depending on the baseline, ANCOVA must be used. In nonrandomized studies of preexisting groups, ANOVA of change seems less biased than ANCOVA, but two control groups and two baseline measurements are recommended.

Keywords: ANCOVA; Change from baseline; Nonrandomized studies; Regression to different means; Regression to the mean; Repeated measures

1. Introduction

The effect of a treatment or exposure on a quantitative outcome, like blood pressure or total score on a clinical questionnaire, is usually evaluated with a ''pretest–posttest control group design.'' The outcome is measured before (pretest, baseline) and after (posttest, outcome) treatment in the treated group and in a control group. Usually, treatment assignment is based on (1) randomization, or (2) baseline values, or (3) preexisting communities. The treatment effect is tested by either of two methods for comparing both groups: (1) analysis of covariance (ANCOVA) with the posttest as outcome and pretest as covariate, or (2) analysis of variance (ANOVA) of the change from baseline, defined as posttest minus pretest. Other methods, such as repeated measures and regression analysis, are equivalent to one of these two, as we will see.

There are several publications on the merits and dangers of both methods [1–11], but most researchers use a single method. The aim of this article is to clarify the purposes and limitations of both methods. This is done by writing both methods as a regression model and as a repeated measures model and applying them to a nonrandomized study of psychotherapy. In Section 2, a definition of treatment effect is given, and the role of randomization, control group, and baseline are discussed. Section 3 applies both methods to a nonrandomized study, showing that they may lead to contradictory conclusions. In Section 4, the methods are compared by two regression equations. It is shown which method is best if treatment assignment is based on randomization (Section 5), the baseline (Section 6), or preexisting groups (Section 7). The article ends with practical advice for nonrandomized studies.

2. Treatment effect and the role of control group, pretest, and randomization

Following [2,6,9], the effect of a treatment G (1 = yes, 0 = no) on an outcome Y for person i is defined as the difference Δ_i between that person's outcome under treatment and under no treatment. The treatment effect for a population

* Corresponding author: Tel.: 0031-433882274 or 0031-433884001.
E-mail address: gerard.vbreukelen@stat.unimaas.nl

0895-4356/06/$ – see front matter
doi: 10.1016/j.jclinepi.2006.02.007

is the average Δ in the population of interest. Most treatments are evaluated by a parallel groups design in which half of all persons are treated (the experimental group, $G = 1$) and half are not (control group, $G = 0$). The mean posttest difference between the groups is used to estimate Δ. In doing so, one assumes that, apart from sampling error, the posttest mean of the control group is equal to the posttest mean of the treated group that would have resulted if that group had not been treated. This assumption is warranted if treatment assignment is based on randomization. However, randomization is not always possible. Exposure studies involve preexisting communities. Mass media interventions can only be implemented at the community level. Treatment contamination may occur if persons within the same school or hospital are allocated to different groups.

But if randomization is impossible, then how can we adjust for a baseline group difference to estimate Δ unbiasedly ? Usually, the outcome is observed before treatment (pretest, X) and after treatment (posttest, Y). If the groups differ significantly at pretest, this invalidates their posttest difference as treatment effect estimator. The next step is adjusting the posttest difference such that Δ is estimated unbiasedly. ANCOVA with treatment G as a factor, pretest X as a covariate, and posttest Y as an outcome, is one attempt at adjustment. ANOVA of change from baseline, with G as a factor and change ($Y - X$) as an outcome, is another one. This article compares both methods in terms of power and bias. To prevent misunderstanding, it must be emphasized that if the groups in a nonrandomized study do not differ at pretest, this does not guarantee that the posttest difference unbiasedly estimates Δ. For instance, the groups may differ in age, and this may lead to a posttest difference even if the pretest means are equal and there is no treatment. A more dramatic example is given in Section 7. For an unbiased effect estimation in nonrandomized studies ‘‘strongly ignorable treatment assignment’’ [8] is needed. Roughly, this means that the actual treatment assignment of person i is independent of Δ_i, and rules out selection of the treatment by each person. Strongly ignorable treatment assignment may hold after correcting for some covariate, which is then called a ‘‘complete confounding factor’’ [11]. For an example, see Section 6. One other assumption is needed to test Δ, that is, ‘‘stable unit-treatment value’’ [8], which comes down to independence between the Δ_i of person i and the treatment assigned to other persons, and rules out treatment contamination.

3. Example: A nonrandomized study of prevention of depression

Before going into the differences between ANCOVA and ANOVA of change, both methods will be applied to a nonrandomized study of prevention of depression [12], which serves as an example throughout this article. The study aim was to evaluate the effectiveness of a psychotherapeutic course in preventing depression among adolescents. The treated group consisted of 88 students, 14–20 years old, in the medium-sized Dutch town Nijmegen, the control group of 92 students in the equally large neighboring town Arnhem. Assessment of symptoms of depression and skills was done before and after intervention. Persons were included if their pretest Beck’s Depression Inventory (BDI) score was between 10 and 25, reflecting mild to moderate depression.

Of all 180 students, 32 dropped out before the posttest: 20 treated and 12 controls. Logistic regression of dropout on treatment, age, gender, schooltype, and pretest of all outcomes, showed dropout to depend on age and schooltype only (all other $P > .30$). The present analysis is limited to complete cases, postponing the inclusion of dropouts until Section 4. So two analyses were run with SPSS: (1) ANOVA of change (post- minus pretest) and (2) ANCOVA with posttest as outcome and pretest as covariate. Both analyses were repeated with age, gender, and schooltype as covariates. Residual checks showed only mild violation of normality and homogeneity of variance. No treatment by pretest interaction was found. Figure 1 shows the result for Symptoms, and plots for skills were similar.

The treated group had a higher pretest mean than the control group ($P = .000$), and the group difference was smaller and no longer significant at posttest (two-tailed $P = .10$). ANOVA of change suggested a treatment effect, because symptoms decreased more in the treated than in the untreated group (effect estimate −0.46, $SE = 0.13$, $P = .000$). In contrast, ANCOVA suggested absence of an effect (estimate −0.13, $SE = 0.13$, $P = 0.31$). These results were hardly affected either by adjusting for covariates or by including dropouts (for details, see Section 4).

4. ANCOVA vs. ANOVA of change: A formal comparison

Traditionally, ANCOVA is treated as an extension of ANOVA [7,13]. Here, we present ANCOVA as a regression

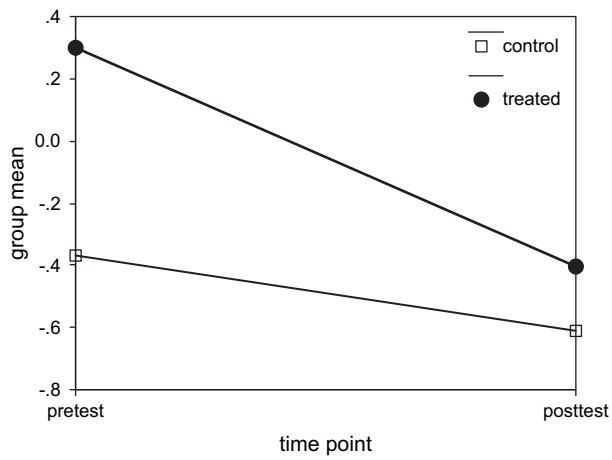

Fig. 1. Change of mean Symptoms score per group in the study of depression prevention.

model, briefly mentioning the difference with classical ANCOVA. In terms of regression analysis, ANCOVA assumes that:

$$Y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2 X_{ij} + e_{ij} \quad (1)$$

or equivalently,

$$\left(Y_{ij} - \beta_2 X_{ij}\right) = \beta_0 + \beta_1 G_{ij} + e_{ij}$$

where Y_{ij} is the posttest score of person i in group j (e.g., $j = 1$ for control, $j = 2$ for treated); G_{ij} is a treatment indicator ($G_{i1} = 0$ for controls, $G_{i2} = 1$ for treated); X_{ij} is the covariate, for example, the pretest score; and e_{ij} is normally distributed with zero mean and constant variance. Classical ANCOVA differs from (1) in that G_{ij} is coded $(-1,+1)$ and X_{ij} is ''centered'' by subtracting its mean [13,14]. This only affects β_0, called the ''grand mean'' in ANOVA.

In eq. (1), β_1 is the group difference on Y adjusted for differences on X. Practical use of ANCOVA requires estimation of β_2, which is a function of the within-group variances and correlation of pretest and posttest. ANCOVA assumes linearity of the covariate effect and absence of covariate by group interaction. Both assumptions can be relaxed [14], but this article is limited to the classical model to allow a comparison with ANOVA of change ($Y - X$), which comes down to (1) with the assumption that $\beta_2 = 1$.

The real difference between ANCOVA and ANOVA of change becomes clear, however, by writing both in terms of repeated measures. ANOVA of change is equivalent to testing the group by time interaction in the following model (with γ instead of β for regression weights to prevent confusion with the ANCOVA model (1):

$$Y_{ijt} = \gamma_0 + \gamma_1 G_{ij} + \gamma_2 T_{it} + \gamma_3 G_{ij} T_{it} + e_{ijt} \quad (2)$$

where Y_{ijt} is the observation of person i in group j at time point t, G is the group (0 = control, 1 = treated), T is the time point (0 = pretest, 1 = posttest), and e_{ijt} is a random person by time effect. Filling in G and T shows that γ_0 is the pretest (population) mean of the control group, γ_1 is the pretest mean difference between the groups, γ_2 is the mean change in the control group, and γ_3 is the difference in mean change between the groups. So testing absence of group by time interaction, that is, of H_0: $\gamma_3 = 0$ in eq. (2), is equivalent to testing the H_0 of no group effect on the change ($Y - X$). Repeated-measures ANOVA differs from (2) only in that it uses $(-1,+1)$ instead of (0,1) coding for G and T.

It is much less known that ANCOVA is equivalent to testing the group by time interaction γ_3 in the reduced model (2), which is obtained by assuming that $\gamma_1 = 0$. So ANCOVA assumes that there is no group difference at pretest [15]. This assumption is warranted if treatment assignment is based on randomization or on the pretest X. In both cases, there is only one group of persons, and so there can be no group effect at pretest. Groups come into existence after the pretest, by randomization or treatment assignment based on X. For these two designs, ANCOVA is known to be the best method [2,9,10].

Repeated-measures analysis of the psychotherapy example in Section 3 was run with the SPSS procedure Mixed, using model (2) with and without the pretest group effect $\gamma_1 G_{ij}$. These two models gave the same effect estimate, *SE* and *P*-value as ANOVA of change and ANCOVA, respectively, confirming that γ_3 in (2) is equivalent to β_1 in (1). An advantage of the repeated measures approach is that it allows inclusion of persons with a missing posttest due to dropout. In this example, including dropouts hardly affected the results. In general, it may make a difference (see Section 7).

In summary, in terms of regression (1), ANOVA of change is a special case of ANCOVA in that it assumes a slope $\beta_2 = 1$ for regressing posttest Y on pretest X. In terms of repeated measures (2), ANCOVA is a special case of ANOVA of change in that it assumes a slope $\gamma_1 = 0$ for regressing pretest X on group G. It is this difference that makes ANCOVA superior in randomized studies and questionable in nonrandomized ones.

5. Randomized studies: Power

In a randomized study any pretest group difference is due to sampling error, so any value of β_2 in (1) gives the same β_1 ($= \Delta$) apart from sampling error, because β_1 is the posttest difference minus $\beta_2 \times$ the pretest difference. ANOVA of the posttest lets $\beta_2 = 0$, ANOVA of change takes $\beta_2 = 1$, and ANCOVA computes β_2 such that the residual posttest variance is minimized, thereby minimizing the standard error of the treatment effect estimate. So ANCOVA gives the largest power and the smallest confidence interval. If pretest and posttest have the same within-group variance and ρ_{XY} denotes the pretest–posttest correlation within groups, then ANCOVA needs a sample size only $(1 + \rho_{XY})/2$ as large as that for ANOVA of change to have the same standard error, for instance, only 75% if $\rho_{XY} = 0.50$.

In terms of repeated measures (2), the superiority of ANCOVA in randomized studies is due to the fact that, because there is no group effect at pretest, ANCOVA is more parsimonious than ANOVA of change, which contains a superfluous parameter γ_1.

In nonrandomized studies the group indicator G in (1) correlates with the pretest X, thereby inflating the SE of the ANCOVA estimator [14]. This explains why both methods gave the same SE in Section 3. But in nonrandomized studies bias, not power, is the issue.

6. Treatment assignment based on the pretest: Bias

Suppose that treatment assignment is based on the pretest X such that the groups have different pretest means.

An example is randomized assignment where the probability of assignment to the treated group increases with X because a high X indicates a strong need for treatment. An extreme case is the "regression discontinuity design" [3], where all persons with X above some cutoff are treated and all persons below it are controls. In these cases the methods cannot both be unbiased, because β_1 in (1) is the posttest difference minus $\beta_2 \times$ the pretest difference. So β_1 depends on β_2 , which is 1 for ANOVA of change, but less than 1 for ANCOVA unless posttest variance is much larger than pretest variance. If pre- and posttest have the same within-group variance, then $\beta_2 = \rho_{XY}$, the within-group correlation. With treatment assignment based on X, ANOVA of change is biased due to regression to the mean while ANCOVA is unbiased [1–3,10,11]. Stratified on X there is random assignment, and so by including X as a covariate the treatment effect Δ is estimated unbiasedly. Compared with pure randomization power is lost, as G in (1) correlates with X.

Regression to the mean may best be understood by taking the case of one group, with pretest and posttest having the same variance. Regression of Y on X then simplifies into: (predicted $Y - \mu_Y$) = $\rho_{XY} \times$ (observed $X - \mu_X$), where ρ_{XY} is the within-group correlation of X and Y, which is less than 1. So the predicted posttest Y is closer to its mean than the observed pretest X used as predictor [16], hence "regression to the mean." This effect may also be understood by noting that if pretest and posttest have the same variance, then it can be shown mathematically that change ($Y - X$) correlates negatively with pretest X. In particular, if the mean change is zero, then high pretest values are on average followed by a decrease, and low pretest values are on average followed by an increase.

That this is not just a mathematical tric, but a real-life phenomenon, can be shown with a simple example. The pretest X of $N = 20$ persons on a symptoms checklist varies from 1 (healthy) to 20 (unhealthy), each person having a different score, which gives a mean of 10.5 and SD of 5.9. A clinician decides to give all persons with $X >$ 10.5 treatment and use all other persons as controls. Unknown to the clinician, no treatment is given at all. Posttest Y is 1 year later, giving again a mean of 10.5 and an SD of 5.9, and a pre–post correlation of 0.52. Figure 2 plots Y against X. Of all 10 persons allocated to treatment (those with $X >$ mean), 6 are below the line $Y = X$. Of all 10 controls (with $X <$ mean), 8 are above the line $Y = X$. So Y is closer to the mean than X for 14 of 20 persons and the posttest group difference is only 4.6 against a pretest difference of 10. ANOVA of change ignores regression to the mean and takes the pretest difference too seriously by subtracting this whole difference from the posttest difference, giving a treatment effect of -5.4 ($P = .03$), where no treatment was given at all. ANCOVA takes regression to the mean into account and subtracts only part of the pretest difference from the posttest difference, leading to the correct conclusion of no effect ($P = 0.60$).

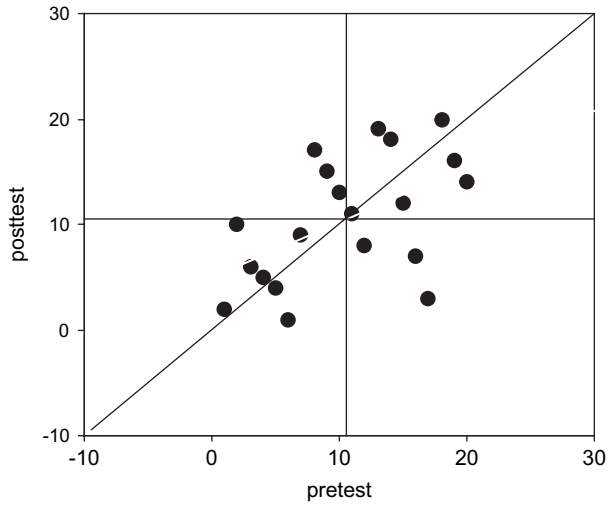

Fig. 2. Regression to the mean effect in the regression discontinuity design. If $X >$ mean, then $Y < X$ (downward regression), and if $X <$ mean, then $Y > X$ (upward regression), for the majority. Reference lines: $X =$ mean (10.5), $Y =$ mean (10.5), $Y = X$.

7. Treatment assigment of preexisting groups: Bias

In nonrandomized studies of preexisting groups these groups often have different pretest means. So β_1 in (1) depends on β_2 and ANOVA of change and ANCOVA cannot both be unbiased and may give contradictory results, a phenomenon known as Lord's ANCOVA paradox [2,4,5]. The difference between the methods can again be shown by the case of no treatment so that $\Delta = 0$, and therefore, $\beta_1 = 0$ must hold for (1) to be unbiased. For ANOVA of change to be unbiased, filling in $\beta_2 = 1$ in (1) shows that the posttest group difference must equal the pretest difference, apart from sampling error. In contrast, ANCOVA gives $\beta_2 < 1$ and so $\beta_1 = 0$ can hold only if the posttest group difference is smaller than the pretest difference. So, whereas ANOVA of change predicts equal change, ANCOVA predicts convergence between groups if there is no treatment.

The reason for this behavior of ANCOVA is its assumption of no group effect at pretest [$\gamma_1 = 0$ in (2)], which leads to regression of both groups to a common mean. If treatment assignment is based on randomization or on the pretest, this assumption is valid, because at pretest no assignment has yet been made and there is only one group. But in a nonrandomized study with preexisting groups it is not obvious toward what population mean the individuals of a group regress [11]. If the two groups are random samples from their populations, and if these populations have different means, then regression of individual scores to the mean of their own population will not change group means apart from sampling error. As a result, the posttest difference equals the pretest difference and ANOVA of change rather than ANCOVA is unbiased, at least if each population has a stable mean or if this mean changes in the same way in both populations. If the two groups are nonrandom samples, then both methods may be biased.

That ANCOVA is biased for preexisting groups, which are random samples from their populations is shown in Fig. 3. The sample of Fig. 2 is now the control group (solid circles) and the experimental group is obtained by adding +10 to each X and Y in the control group (clear circles). So the group difference is 10 at both time points. There is no treatment in either group and so $\Delta = 0$. ANOVA of change correctly estimates β_1 to be zero ($P = 1.00$), but ANCOVA estimates β_1 to be 4.8 ($P = .03$). Almost the same result (effect = 5.4, $P = .03$) is obtained by first matching on X, which leads to the exclusion of all persons with $X < 11$ or $X > 20$ (vertical lines in Fig. 3), and then applying ANOVA to the posttest Y or the change $Y - X$ of included persons only. Figure 3 shows the cause of this bias. Matching leads to selection of the upper half of control group I and the lower half of experimental group II. At posttest there is regression, not to a common mean as ANCOVA assumes, but to the mean that would have been observed without selection, that is, to 10.5 in the control group and 20.5 in the experimental group. Of all 10 included controls, 6 are below the $Y = X$ line (downward regression). Of all 10 included experimentals, 8 are above it (upward regression). The opposite trends occur in the excluded subgroups. ANCOVA is a mathematical method of matching and shares its bias in nonrandomized studies.

In this example the bias is clear, because there is no treatment and we have posttest data of all persons. In practice, there are no posttest data of excluded persons and ANOVA of change on the included (matched) persons suffers from the same differential regression effect as ANCOVA on the total sample. This is a problem in nonrandomized studies where the inclusion of persons is based on cutoffs for the pretest, which act like a mild matching. But there is a simple solution. If exclusion is based on the pretest data, then posttest data of the excluded persons are ''missing at random'' [17]. This type of missingness can be handled by repeated-measures analysis (2), including the pretest data of excluded persons, which is not possible with (1). In the present example repeated-measures analysis of all 40 persons, using only the pretest data of excluded persons, gave an effect estimate of 2.55, with a two-tailed $P = .30$, leading to the correct conclusion of no effect, while ANCOVA of all data and ANOVA of change without excluded persons led to the wrong conclusion.

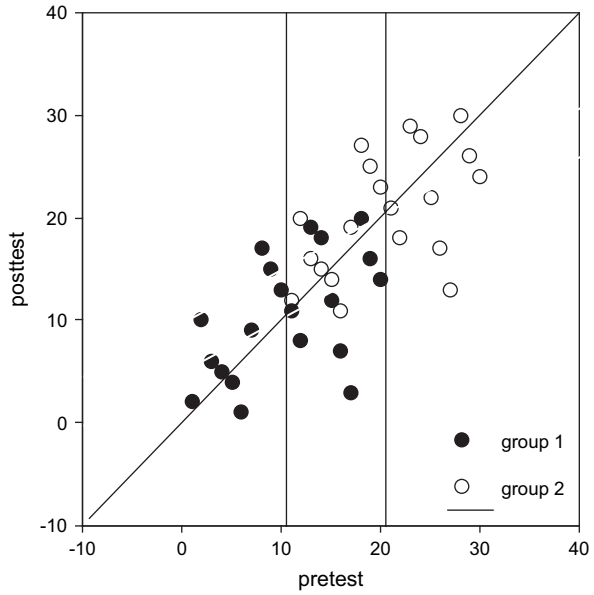

Fig. 3. Bias introduced by matching on the pretest X due to regression to different means, in a nonrandomized study of preexisting groups with a fixed mean each. Reference lines: $X = 10.5$ and $X = 20.5$ (inclusion criterion: $10 < X < 21$) and $Y = X$. ●: Included: $X > 10$, result: $Y < X$; excluded: $X < 10$, result: $Y > X$, for a majority. ○:Included: $X < 20$, result: $Y > X$; excluded: $X > 20$, result: $Y < X$, for a majority.

What do these results imply for the example in Section 3? Given that the groups were recruited from different towns and had different pretest means, ANOVA of change seems more reasonable than ANCOVA, and one might conclude that there was a treatment effect. But there are two complications. First, the BDI score was used both as inclusion criterion ($10 < \text{BDI} < 25$) and as part of the outcome Symptoms, implying some matching on the pretest. As Fig. 3 shows, this may lead to differential regression if the two populations (before exclusion based on BDI) have different BDI means. This not only threatens the unbiasedness of ANCOVA, but also that of ANOVA of change when applied to the included persons only. But because no data are available from the excluded persons, no further analysis is possible. A second complication is that the BDI score was measured twice before the intervention period in the control group. The first was used as inclusion criterion and the second was 1 month later on the pretest of all outcomes ([12], p. 142). Given that a BDI score >10 is well above the population mean ([12], p. 72), it is likely that regression to the mean had already occurred at pretest in the control group. So the pretest difference in Fig. 1 may be artificially large, casting doubt on the treatment effect. Unfortunately, no data from that first BDI measurement in the control group are available.

8. Discussion

Based on literature, we saw that (1) the difference between ANCOVA and ANOVA of change is that between assuming absence or presence of a baseline group difference, and (2) the choice between both methods depends on the treatment assignment procedure. If treatment assignment is by randomization, both methods are unbiased but ANCOVA has more power. If treatment assignment is based on the pretest, ANCOVA is unbiased but ANOVA of change is not, due to regression to the mean. Both designs imply treatment assignment after the pretest and so at pretest there is one group, justifying the ANCOVA assumption of no group effect at pretest. In contrast, if preexisting groups are assigned to treatment, the unbiasedness of both methods

depends on strong assumptions about trends in the absence of treatment. The ANOVA of change assumption (equal change) is more plausible than the ANCOVA assumption (regression to a common mean), at least if each group is a random sample from its population. The larger the pretest difference between preexisting groups, the worse ANCOVA is on bias (Section 7) and efficiency (Section 5). This bias has to do with measurement error (intraindividual variability) in the covariate, which leads to underestimation of β_2 in (1) and a greater discrepancy between ANCOVA and ANOVA of change. Statistical corrections for this underestimation exist [7,18], but are beyond the present scope. Instead, measurement error can be reduced by repeated pretesting and taking a person's average as covariate [19].

The present results lead to practical advice for nonrandomized studies, assuming that person and cluster randomization are impossible. The design is enhanced by having (1) more than one control group [20], and (2) more than one pretest [7,11], and (3) more than one outcome, including some that are known to be unaffected by treatment [21]. Having more than one control group or pretest allows estimation of group trend in the absence of treatment. Equal change of two control groups provides support for ANOVA of change, especially if the treated group is in-between both control groups at pretest. Likewise, equal change between repeated pretests of treated and control group suggests the use of ANOVA of change rather than ANCOVA. Moreover, a person's average pretest is less subject to measurement error than a single pretest, leading to larger power for both methods and less disagreement. Finally, including outcomes known to be unaffected by treatment allows checks on hidden bias in methods of analysis. For instance, finding a treatment effect on intelligence in the study of depression would cast doubt on the method of analysis. In nonrandomized studies with one preexisting control group and one pretest, ANOVA of change may be better than ANCOVA, but running both methods may be even better. If both methods lead to the same conclusion, differing only in effect size, this increases one's confidence in that conclusion [3].

Additional problems, illustrated by the study of depression, are dropout and bias due to inclusion criteria in nonrandomized studies. In randomized and nonrandomized studies, dropouts must be included by using proper methods for missing data [17]. In nonrandomized studies, further bias arises from the exclusion of persons based on pretest data, as Fig. 3 showed. So the best analysis of preexisting groups may be a repeated-measures analysis including all available data of dropouts and of excluded persons.

References

[1] Campbell DT, Kenny DA. A primer on regression artifacts. New York: Guilford Press; 1999.

[2] Holland PW, Rubin DB. On Lord's paradox. In: Wainer H, Messick S, eds. Principals of modern psychological measurement. Hillsdale, NJ: Erlbaum; 1983. p. 3–25.

[3] Kenny DA. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychol Bull 1975;82:345–62.

[4] Lord FM. A paradox in the interpretation of group comparisons. Psychol Bull 1967;68:304–5.

[5] Lord FM. Statistical adjustments when comparing pre-existing groups. Psychol Bull 1969;72:336–7.

[6] Maris E. Covariance adjustment versus gain scores revisited. Psychol Methods 1998;3:309–27.

[7] Porter AC, Raudenbush SW. Analysis of covariance: its model and use in psychological research. J Counseling Psychol 1987;34:383–92.

[8] Rosenbaum PR, Rubin DB. Estimating the effects caused by treatments. J Am Stat Assoc 1984;79:26–8.

[9] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Ed Psychol 1974;66:688–701.

[10] Rubin DB. Assignment to treatment group on the basis of a covariate. J Ed Stat 1977;2:1–26.

[11] Weisberg HI. Statistical adjustments and uncontrolled studies. Psychol Bull 1979;86:1149–64.

[12] Ruiter M. Preventie van depressie bij jongeren (Prevention of depression among adolescents). Doctoral dissertation, Nijmegen University, The Netherlands; 1997.

[13] Maxwell SE, Delaney HD. Designing experiments and analyzing data: a model comparison perspective. Pacific Grove, CA: Brooks/Cole; 1990.

[14] Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and other multivariable methods Pacific Grove, CA: Brooks/Cole; 1998.

[15] Laird NM, Wang F. Estimating rates of change in randomized clinical trials. Controlled Clin Trials 1990;11:405–19.

[16] Stigler SM. Regression towards the mean, historically considered. Stati Methods Med Res 1997;6:103–14.

[17] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.

[18] Carroll RJ. Covariance analysis in general linear measurement error models. Stat Med 1989;8:1075–93.

[19] Senn SJ. Covariance analysis in generalized linear measurement error models. Stat Med 1990;9:583–6.

[20] Rosenbaum PR. The role of a second control group in an observational study. Stat Sci 1987;2:292–316.

[21] Rosenbaum PR. The role of known effects in observational studies. Biometrics 1987;45:557–69.