Research article

The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study

Andrew J Vickers

Address: Integrative Medicine Service, Biostatistics Service, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue New York, New York 10021, USA

E-mail: vickersa@mskcc.org

Published: 28 June 2001

BMC Medical Research Methodology 2001, **1**:6

Received: 28 March 2001
Accepted: 28 June 2001

This article is available from: http://www.biomedcentral.com/1471-2288/1/6

Abstract

Background: Many randomized trials involve measuring a continuous outcome - such as pain, body weight or blood pressure - at baseline and after treatment. In this paper, I compare four possibilities for how such trials can be analyzed: post-treatment; change between baseline and post-treatment; percentage change between baseline and post-treatment and analysis of covariance (ANCOVA) with baseline score as a covariate. The statistical power of each method was determined for a hypothetical randomized trial under a range of correlations between baseline and post-treatment scores.

Results: ANCOVA has the highest statistical power. Change from baseline has acceptable power when correlation between baseline and post-treatment scores is high;when correlation is low, analyzing only post-treatment scores has reasonable power. Percentage change from baseline has the lowest statistical power and was highly sensitive to changes in variance. Theoretical considerations suggest that percentage change from baseline will also fail to protect from bias in the case of baseline imbalance and will lead to an excess of trials with non-normally distributed outcome data.

Conclusions: Percentage change from baseline should not be used in statistical analysis. Trialists wishing to report this statistic should use another method, such as ANCOVA, and convert the results to a percentage change by using mean baseline scores.

Background

Many randomized trials involve measuring a continuous outcome at baseline and after treatment. Typical examples include trials of pravastatin for hypercholesterolemia [1], exercise and diet for obesity in osteoarthritis patients [2] and acupuncture for pain in athletes with shoulder injuries [3]. In each trial, the outcome measure used to determine the effectiveness of treatment - cholesterol, body weight or shoulder pain - was measured both before treatment had started and after it was complete.

In the case of a single post treatment outcome assessment, there are four possibilities for how such data can be entered into the statistical analysis of such trials. One can use the baseline score solely to ensure baseline comparability and enter only the post-treatment score into analysis (I will describe this method as "POST"). Alternatively, one can analyze the change from baseline, either by looking at absolute differences ("CHANGE") or a percentage change from baseline ("FRACTION"). The most sophisticated method is to construct a regression model which adjusts the post-treatment score by the baseline

If b is the score at baseline, f is the score after treatment and g is treatment group:

POST: Compare f by g
FRACTION: Compare $\frac{b\text{-}f}{b} x100$ by g
CHANGE: Compare $b\text{-}f$ by g
ANCOVA: Calculate $f = \text{constant} + \beta_1 b + \beta_2 g$ and report β_2

Figure 1
Mathematical description of the four methods

POST: Pain score was 10 points lower in the treatment group
FRACTION: Pain fell by 40% in the treatment group but only 10% in controls
CHANGE :Pain fell by 15 points more in the treatment group
ANCOVA: After adjusting for baseline differences, pain fell by 15 points more in the treatment group.

Figure 2
Examples of the results of a trial analyzed by each method in ordinary language terms

score ("ANCOVA"). Figure 1 describes each of these methods in mathematical terms. Figure 2 gives examples of the results of each method described in ordinary language.

Some trials assess outcome several times after treatment, a design known as "repeated measures." Each of the four methods described above can be used to analyze such trials by using a summary statistic such as a mean or an "area-under-curve" [4]. There are several more complex methods of analyzing such data including repeated measures analysis of variance and generalized linear estimation [5]. These methods are of particular value when the post-treatment scores have a predictable course over time (e.g. quality of life in late stage cancer patients) or when it is important to assess interactions between treatment and time (e.g. long-term symptomatic medication). This paper will concentrate on the simpler case where time in not an important independent variable.

The choice of which method to use can be determined by analysis of the statistical properties of each. An important criteria for a good statistical method is that it should reduce the rate of false negatives (β). The β of a statistical test is usually expressed in terms of statistical power (1-β). Power is normally fixed, typically at 0.8 or 0.9, and the required amount of data (e.g. number of evaluable patients) is calculated. A method that requires relatively fewer data to provide a certain level of statistical power is described as *efficient*.

The characteristics of the four methods - POST, CHANGE, FRACTION and ANCOVA - have been studied by statisticians for some time [6, 7, 8]. In this paper, I aim to provide statistical data that can guide clinical research yet is readily comprehensible by non-statisticians. Accordingly, I will compare the methods using a hypothetical trial and express results in terms of statistical power.

Methods

All calculations and simulations were conducted using the statistical software Stata 6.0 (Stata Corp., College Station, Texas). I created a hypothetical pain trial with patients divided evenly between a treatment and a control group. The pain score for any individual patient was sampled from a normal distribution. The mean score at

baseline was 50 mm on a visual analog scale of pain (VAS); after treatment, mean pain was expected to be 50 mm in controls and 45 mm in treated patients. The standard deviation of all scores was 10. The text of the simulation is given in the appendix (appendix.doc).

I calculated the statistical power of the different methods of analysis for this trial given a sample size of 100 patients. As power varies according to the correlation between baseline and follow-up scores, a range of different possible correlations were used. The power for POST, CHANGE and ANCOVA were calculated using the "sampsi" function of Stata. This derives power analytically using formula developed by Frison and Pocock [6]. The power for FRACTION was calculated by the simulation described above. The simulation was first validated against Stata's results for POST and CHANGE at a correlation of 0.5. It was then conducted using 1000 repetitions calculating ttests for FRACTION at a range of correlations between 0.2 and 0.8. The number of results in which p was less than 5% was calculated.

Results and Discussion

The true positive rates of the four statistical methods given different correlations are given in table 1. These data are equivalent to statistical power, or 1-β. As has been previously reported [6], ANCOVA has the highest statistical power. CHANGE has acceptable power when correlation between baseline and post-treatment scores are high;when correlations are low, POST has reasonable power. FRACTION has poor statistical efficiency at all correlations.

Table 1: Statistical power of each method of analysis

Correlation	ρ = 0.2	ρ = 0.35	ρ = 0.5	ρ = 0.65	ρ = 0.8
POST	70.5%	70.5%	70.5%	70.5%	70.5%
FRACTION	45.1%	56.4%	67.0%	82.7%	97.1%
CHANGE	50.7%	59.2%	70.5%	84.8%	97.7%
ANCOVA	72.3%	76.1%	82.3%	90.8%	98.6%

Moreover, the power of FRACTION is sensitive to changes in the characteristics of the baseline distribution. If the range of baseline values is large, the variance of FRACTION increases disproportionately and power falls. Simulations were repeated with the standard deviations and difference between groups doubled. There was no difference in the power of POST, CHANGE or ANCOVA. The power of FRACTION fell dramatically: at correlations of 0.2, 0.35, 0.5, 0.65 and 0.8 respectively, power was 18%, 24%, 33%, 45% and 63%.

It is arguable that the method of simulation is biased against FRACTION because the treatment effect is additive, that is, the simulation models an absolute 5 mm difference between groups. In theory, the difference between FRACTION and CHANGE should decrease if the treatment effect is proportional. The simulation was therefore repeated with the treatment group experiencing an average 10% decrease from baseline. Correlation between baseline and follow-up scores was varied randomly between 0.2 and 0.8. The p values from a ttest of FRACTION and CHANGE were directly compared over 1000 simulations: p values were lower for CHANGE approximately 65% of the time.

Theoretical considerations suggestion two further disadvantages to FRACTION. First, because it incorporates both baseline and post-treatment scores, it would appear to control for any chance baseline imbalance between groups. However, this is not the case because of regression to the mean: FRACTION will create a bias towards the group with poorer baseline scores (the same is true for CHANGE; POST causes bias in the opposite direction). Second, because it is calculated using a ratio, it may cause outcome data to be non-normally distributed. In a bivariate normal distribution (such as a baseline and post-treatment score) any statistic using either variable alone or combining both by addition or subtraction will be normally distributed. There is no analytic reason why a statistic created by multiplying or dividing one variable by the other should necessarily have a normal distribution.

Conclusion

Reporting a percentage change from baseline gives the results of a randomized trial in clinically relevant terms immediately accessible to patients and clinicians alike. This is presumably why researchers investigating issues such as the effects of medication on hot flashes [9], or of different chemotherapy regimes on quality of life [10], report this statistic.

However, percentage change from baseline is statistically inefficient. Perhaps counterintuitively, it does not correct for imbalance between groups at baseline. It may also create a non-normally distributed statistic from normally distributed data. Percentage change from baseline should therefore not be used in statistical analysis. Trialists wishing to report percentage change should first use another method, preferably ANCOVA, to test significance and calculate confidence intervals. They should then convert results to percentage change by using mean

baseline and post-treatment scores. For an example of this approach, see Crouse et al. [11].

The findings presented here reconfirm previously reported data suggesting that ANCOVA is the method of choice for analyzing the results of trials with baseline and post treatment measurement. In cases where ANCOVA cannot be used, such as with small samples or where the assumptions underlying ANCOVA modeling do not hold, CHANGE or POST are acceptable alternatives, especially baseline variables are comparable between groups (perhaps ensured by stratification) and if correlation between baseline and post-treatment scores are either high (for CHANGE) or low (for POST). The use of FRACTION should be avoided.

Competing interests

Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this paper?

No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?

No

Do you have any other financial competing interests?

No

Are there any non-financial competing interests you would like to declare in relation to this paper?

No

Additional material

Text of simulation
[http://www.biomedcentral.com/content/supplementary/1471-2288-1-6-S1.doc]

References

1. Glorioso N, Troffa C, Filigheddu F, Dettori F, Soro A, Parpaglia PP, Collatina S, Pahor M: **Effect of the HMG-CoA reductase inhibitors on blood pressure in patients with essential hypertension and primary hypercholesterolemia.** *Hypertension* 1999, **34**:1281-1286
2. Messier SP, Loeser RF, Mitchell MN, Valle G, Morgan TP, Rejeski WJ, Ettinger WH: **Exercise and weight loss in obese older adults with knee osteoarthritis: a preliminary study.** *J Am Geriatr Soc* 2000, **48**:1062-1072
3. Kleinhenz J, Streitberger K, Windeler J, Gussbacher A, Mavridis G, Martin E: **Randomised clinical trial comparing the effects of acupuncture and a newly designed placebo needle in rotator cuff tendinitis.** *Pain* 1999, **83**:235-241
4. Matthews JNS, Altman D, Campbell MJ, Royston P: **Analysis of serial measurements in medical research.** *BMJ* 1990, **300**:230-235
5. Omar RZ, Wright EM, Turner RM, Thompson SG: **Analysing repeated measurements data: a practical comparison of methods.** *Stat Med* 1999, **18**:1587-1603
6. Harris CW Ed: **Problems in measuring change.** *Madison: University of Wisconsin Press* 1967
7. Frison L, Pocock SJ: **Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design.** *Stat Med* 1992, **11**:1685-1704
8. S Senn: **Statistical Issues in Drug Development.** *Chichester: John Wiley* 1997
9. Loprinzi CL, Michalak JC, Quella SK, O'Fallon JR, Hatfield AK, Nelimark RA, Dose AM, Fischer T, Johnson C, Klatt NE, *et al*: **Megestrol acetate for the prevention of hot flashes.** *N Engl J Med* 1994, **331**:347-352
10. Anderson H, Hopwood P, Stephens RJ, Thatcher N, Cottier B, Nicholson M, Milroy R, Maughan TS, Falk SJ, Bond MG, Burt PA, Connolly CK, McIllmurray MB, Carmichael J: **Gemcitabine plus best supportive care (BSC) vs BSC in inoperable non-small cell lung cancer - a randomized trial with quality of life as the primary outcome. UK NSCLC Gemcitabine Group. Non-Small Cell Lung Cancer.** *Br J Cancer* 2000, **83**:447-453
11. Crouse JR 3rd, Morgan T, Terry JG, Ellis J, Vitolins M, Burke GL: **A randomized trial comparing the effect of casein with that of soy protein containing varying amounts of isoflavones on plasma concentrations of lipids and lipoproteins.** *Arch Intern Med* 1999, **159**:A2070-2076

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/content/backmatter/1471-2288-1-6-b1.pdf