Misleading Comparisons: The Fallacy of Comparing Statistical Significance

Kristin Sainani, PhD

This article will review a common statistical fallacy: that it is possible to make conclusions by comparing statistical significance or *P* values (rather than effect sizes) between groups. For example, if a treatment group experiences a statistically significant improvement in outcomes, whereas a control group does not, it may be tempting to conclude that treatment is superior to control—but this conclusion does not follow. This article will review how this fallacy arises in 2 contexts and will give readers tips on how to avoid being misled.

EXAMPLE 1: BETWEEN-GROUP DIFFERENCES VERSUS WITHIN-GROUP DIFFERENCES

When a study contains a control group, the comparison of interest is whether the treatment group has better (or worse) outcomes than the control group. Readers should be wary when authors omit or gloss over such **between-group differences** and instead focus on **within-group differences**. For example, authors may report that "the treatment group improved significantly from baseline, but the control group did not." This result tells the reader nothing about how the treatment group fared relative to the control group—and readers should not be fooled into thinking that it does.

Between-group difference: assesses whether groups differ significantly; for example, whether the change in one group is bigger than the change in another group.

For example, consider a randomized, double-blind, controlled trial of the omega-3 fatty acid docosahexaenoic acid (DHA) to treat eczema [1], a common skin condition. The control group received a non-DHA fatty acid, and improvements in eczema symptoms were evaluated using the SCORAD (severity score of atopic dermatitis) index. The study reports the following results in the abstract: "DHA, but not the control treatment, resulted in a significant clinical improvement of atopic eczema in terms of a decreased SCORAD." Only much later, in the discussion section, is it revealed that: "There was no significant difference between the DHA and control groups."

Within-group difference: assesses whether the change in one group is significantly different than 0.

By burying these between-group findings late in the text of the paper, the authors give readers the false impression that the study results were positive, when in fact the results were null. Indeed, the study garnered considerable media attention—with all reports mistakenly implying that the DHA had shown positive effects in a randomized, controlled trial (for examples, just Google "DHA and eczema").

Wilcoxon sign-rank test: a statistical test for assessing within-group chances when the outcome variable is not normally distributed (such as SCORAD).

Figure 1 graphically displays the results from the trial. The DHA group had a median decline of 18% in SCORAD ($P = .0009$, **Wilcoxon sign-rank test**), whereas the control group had a median decline of 11%, which did not reach statistical significance. The 7% difference in improvement between the groups was not statistically significantly (an important piece of information omitted from the graph). Although it is true that the improvements in the DHA group were greater than in the control group, this difference could be due to chance. Thus, the improvements in the DHA group may be attributable to spontaneous resolution of symptoms, supplementation with any fatty acid, a placebo effect, or a combination of these (which the authors do acknowledge in their discussion).

K.S. Department of Health Research and Policy, Stanford University, Stanford, CA 94305. Address correspondence to: K.S.; e-mail: kcobb@stanford.edu

Disclosure: nothing to disclose

Disclosure Key can be found on the Table of Contents and at www.pmrjournal.org

Submitted for publication April 19, 2010; accepted April 19, 2010.

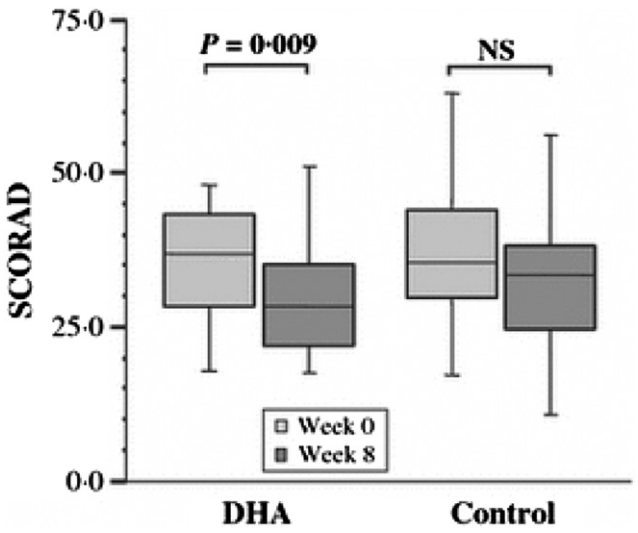

Figure 1. Results from a randomized controlled trial of docosahexaenoic acid (DHA) to treat eczema. Both the DHA and control groups improved. Though the within-group changes were only significant for the DHA group, the improvement in the DHA group (a median 18% decline) was not statistically different than the improvement in the control group (a median 11% decline). Reproduced with permission from Wiley (1).

It seems surprising that 2 groups could differ with regard to within-group significance and yet not differ significantly from one another. But, in fact, this is a common occurrence. Table 1 shows 4 hypothetical examples of how this can occur. Statistical significance depends on the following factors: effect size, standard deviation (variability), and sample size. Thus, within-group statistical significance may differ between groups because of any of these factors (or a combination of these factors)—not just effect sizes. In the DHA trial, both a smaller effect size and greater variability in SCORAD contributed to a lack of within-group significance in the control group. Additionally, for the same effect size, it is easier to achieve within-group significance than between-group significance (because 2-sample statistics have to contend with variability from 2 sources; see In-Depth Box for more details).

When considering controlled studies, readers should determine whether reported *P* values refer to within-group or between-group comparisons. The choice of statistical test is an important clue—for example, the paired *t*-test is used for within-group comparisons, whereas the 2-sample *t*-test is used for between-group comparisons (see Table 2 for more examples). Readers should be skeptical when authors have emphasized within-group changes above between-group changes. If the authors fail to report comparisons against the control group (or these comparisons are buried in the paper), this is likely an example where the authors have tried to "spin" a null result.

Table 2. *Examples of statistical tests used to evaluate within-group effects versus statistical tests used to evaluate between-group effects*

Statistical tests for within-group effects	Statistical tests for between-group effects
Paired *t*-test	Two-sample *t*-test
Wilcoxon sign-rank test	Wilcoxon sum-rank test (equivalently, Mann-Whitney U test)
Repeated-measures ANOVA, time effect	ANOVA; repeated-measures ANOVA, group*time effect
McNemar test	Difference in proportions, chi-square test, or relative risk

ANOVA = analysis of variance.

EXAMPLE 2: WITHIN-SUBGROUP SIGNIFICANCE VERSUS INTERACTION

A more subtle—but parallel—instance of this fallacy arises in the context of **interaction**. Interaction occurs when the effect of a particular treatment differs in different subgroups. For example, if a drug works significantly better in men than women, we say that there is an interaction between drug and gender, or, equivalently, that the effect of the drug is modified by gender. It is a common misconception that one can prove an interaction by demonstrating that an effect is significant in 1 subgroup but not another (for example, that the drug significantly improves blood pressure in men but not in women). However, this does not constitute adequate proof of interaction. Rather, one must demonstrate that the effects differ significantly between the groups (ie, that the

Interaction: the effect of a treatment (or exposure) differs significantly between different subgroups. For example, a drug reduces blood pressure significantly more in men than in women.

Table 1. *Four hypothetical examples in which within-group significance differs between 2 groups, but the between-group difference is not significant**

Group 1				Group 2				
Effect size	**Standard deviation**	**Sample size**	**Within-group *P* value**	**Effect size**	**Standard deviation**	**Sample size**	**Within-group *P* value**	**Between-group *P* value**
10	20	**30**	.01	10	20	**10**	.15	1.00
10	**15**	20	.008	10	**30**	20	.15	1.00
10	15	20	.008	**5**	15	20	.15	.30
10	**10**	**20**	.0003	**15**	**30**	**10**	.15	.36

*Within-group *P* values are calculated using paired *t*-tests; between-group *P* values are calculated using 2-sample *t*-tests. Bolded inputs differ between the groups.

IN DEPTH: WITHIN-GROUP COMPARISONS VERSUS BETWEEN-GROUP COMPARISONS.

For the same effect size (ES), sample size per group (n), and standard deviation (SD), it is easier to achieve statistical significance with a within-group comparison than a between-group comparison, as illustrated by comparing the formula for a paired *t-test* with that for a 2-sample *t-test*:

The paired *t-test* is used to determine whether the change within 1 group is statistically significant (for continuous, normally distributed outcomes):

$$T_{n-1} = \frac{ES}{\sqrt{\frac{SD^2}{n}}} = \frac{\sqrt{n}}{SD} ES$$

The 2-sample *t-test* is used to determine whether the difference between 2 groups is statistically significant (for continuous, normally distributed outcomes):

$$T_{2n-2} = \frac{ES}{\sqrt{\frac{SD^2}{n} + \frac{SD^2}{n}}} = \frac{\frac{\sqrt{n}}{SD} ES}{\sqrt{2}}$$

Thus, for the same ES, SD, and n-per-group, the T statistic for the 2-sample *t-test* will always be smaller than the T statistic for the paired *t-test* by a factor of the square root of 2. Smaller T values translate to higher *P* values (except in some cases where n is very small—roughly 4 or less—because of differences in the degrees of freedom between the 2 T statistics).

improvement in blood pressure in men taking the drug is significantly greater than the improvement in blood pressure in women taking the drug).

For illustration, consider a recent 4-group randomized trial for smoking cessation [2]. Weight-concerned women smokers were randomly assigned to weight-focused counseling or standard counseling plus the drug bupropion or a placebo pill. The primary outcomes were prolonged abstinence rates at 3, 6, and 12 months. Table 3 shows the results of the trial. The effect of adding bupropion is somewhat larger in the weight-focused counseling group—for example, bupropion boosted 1-year abstinence rates by 16% in this group versus 12% in the standard counseling group. The effect of bupropion is also statistically significant in the weight-focused counseling group (*P* values: .001, .001, .006), whereas it just misses statistical significance in the standard counseling group (*P* values: .07, .08, .05). However, it does not follow from this that the effect of bupropion differs by counseling type. The authors did not report formal tests for interaction, which directly compare the effects of bupropion in the 2 counseling groups (eg, 16% improvement in 1-year abstinence rates versus 12% improvement). However, I ran these tests based on data available in the paper and found no evidence of interaction (Table 3). Thus, although the effects of bupropion are larger in the weight-focused counseling group, these differences could easily be due to chance. Note also that the effect of bupropion in the standard counseling group would have achieved statistical significance had the sample size of this group been as large as that of the weight-focused counseling group.

The authors are appropriately cautious in their conclusions and never claim to have found an interaction. But, their final conclusion that: "among weight-concerned women smokers, bupropion therapy increased cessation rates when added to a specialized weight concerns intervention, but not when added to standard counseling" may give readers the

Table 3. *Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a 4-arm randomized trial of smoking cessation**

Months after quit target date	Weight-focused counseling			Standard counseling group			
	Bupropion group abstinence (n = 106)	Placebo group abstinence (n = 87)	*P* value, bupropion vs placebo	Bupropion group abstinence (n = 89)	Placebo group abstinence (n = 67)	*P* value, bupropion vs placebo	*P* value for interaction between bupropion and counseling type†
3	41%	18%	.001	33%	19%	.07	.42
6	34%	11%	.001	21%	10%	.08	.39
12	24%	8%	.006	19%	7%	.05	.79

*From Tables 2 and 3 of Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and cognitive behavioral therapy for weight-concerned women smokers. Arch Intern Med 2010;170:543-550.

†Interaction *P* values were newly calculated from logistic regression based on the abstinence rates and sample sizes shown in this table.

false impression that the effect of bupropion differs by counseling type. What the data actually show is that there is a strong main effect for bupropion (adding bupropion improves abstinence rates when counseling groups are collapsed), there is no main effect for weight-focused counseling (weight-focused counseling does not improve abstinence over standard counseling when bupropion and placebo groups are collapsed), and there is no interaction between bupropion and counseling (the effect of bupropion does not differ statistically between the two counseling groups). So, a better final conclusion would be that bupropion improves abstinence rates over counseling of any type.

CONCLUSION

Readers should be wary of comparisons of statistical significance, such as "group A improved significantly from baseline but group B did not" or "the effect of the treatment was significant in subgroup A but not subgroup B." There is little that can be concluded from these types of comparisons. When presented with such results, readers should: compare the effect sizes in the groups (or subgroups) to determine whether the differences appear clinically important and look for formal between-group comparisons or tests of interaction to determine whether the differences are statistically significant. If these tests are not provided, it may be possible to calculate them based on data available in the paper. Readers should have a particularly high index of suspicion for controlled studies that fail to report between-group comparisons, because these likely represent attempts to "spin" null results.

REFERENCES

1. Koch C, Dölle S, Metzger M, et al. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. Br J Dermatol 2008;158:786-792.
2. Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and cognitive behavioral therapy for weight-concerned women smokers. Arch Intern Med 2010;170:543-550.