PM R 9 (2017) 933-937

www.pmrjournal.org

Statistically Speaking

Getting the Right Answer: Four Statistical Principles

Kristin L. Sainani, PhD

Introduction

Researchers often go into studies with preconceived notions of what they expect or hope to see. These expectations and hopes may color how researchers interpret their final data. For example, they may cherry-pick results that appear to support their hypothesis while ignoring results that contradict it. This can lead to biased papers, with flawed analyses, even including examples in which authors wrongly report the exact opposite of what the data actually show. I will present one such example here. I contend that researchers can avoid such egregious errors if they follow 4 simple statistical principles: (1) plot the data; (2) create focused and uncluttered tables; (3) consider the totality of the evidence; and (4) focus on effect sizes more than *P* values.

Example

In a randomized, double-blind study published in 2016 in *Aging Clinical and Experimental Research*, researchers assigned 50 middle-aged and older adults with knee pain and/or stiffness to a supplement—*N*-acetyl glucosamine and chondroitin sulfate—or a placebo pill for 24 weeks [1]. The 2 primary outcomes were pain and osteoarthritis severity. Four secondary outcomes included a household activity score, a leisure activity score, the Timed Up and Go test, and the 6-Minute Walk Test.

The authors reported that the treatment works, concluding that: "Regularly taking *N*-acetyl glucosamine and chondroitin sulfate supplements may help middle-aged and older Japanese adults with knee pain and/or stiffness obtain a more active lifestyle." However, a closer inspection of the data reveals that this study had a null result—the treatment appears no more effective than placebo.

How could the authors have gotten it wrong? Deliberately or inadvertently, they made several mistakes that led them to the wrong answer, including failing to plot the data, cluttering tables with distracting and irrelevant data, cherry-picking isolated results, and paying too much attention to *P* values and too little attention to effect sizes. I will review their data to show how these mistakes led to the wrong answer.

Plot the Data

The researchers mistakenly concluded that the treatment group had greater reductions than the placebo group in osteoarthritis severity (as measured by the Japanese Knee Osteoarthritis Measure). A simple plot of the data would have revealed their error.

I used data from Table 2 of their study [1] to plot changes in the 2 primary outcomes—pain and osteoarthritis severity—by group (Figure 1). Figure 1 clearly shows a null result. The placebo group actually had a slightly greater decrease in pain than the treatment group, and the 2 groups had near-identical decreases in osteoarthritis severity by 24 weeks.

The authors correctly acknowledged the null result for pain but mistakenly concluded: "These results suggest that consumption of *N*-acetyl glucosamine and chondroitin sulfate for 12 weeks or longer has a positive effect on osteoarthritis score." They arrived at this faulty conclusion by overinterpreting a significant *P* value—the time x group interaction from repeated-measures analysis of variance was significant ($P = .045$) for osteoarthritis score. Although this is the correct *P* value to consider, a significant time x group interaction only tells us that the groups differed somewhere in their patterns of change over time; it does not tell us where those differences lie. In this example, the 2 groups differed as to exactly when osteoarthritis score peaked and dipped (Figure 1B), which triggered the significant time x group *P* value . But, as the plot makes obvious, the 2 groups had near-identical improvements over the course of the trial.

Create Focused and Uncluttered Tables

The authors also got the wrong answer because of poor data presentation. They needlessly cluttered their tables with irrelevant and distracting details. Figure 2

1934-1482/$ - see front matter © 2017 by the American Academy of Physical Medicine and Rehabilitation
http://dx.doi.org/10.1016/j.pmrj.2017.06.015

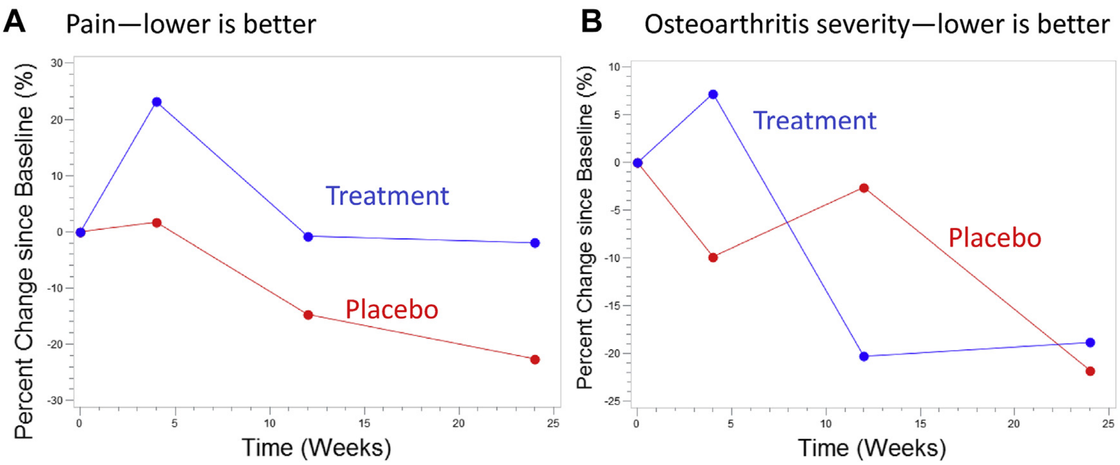
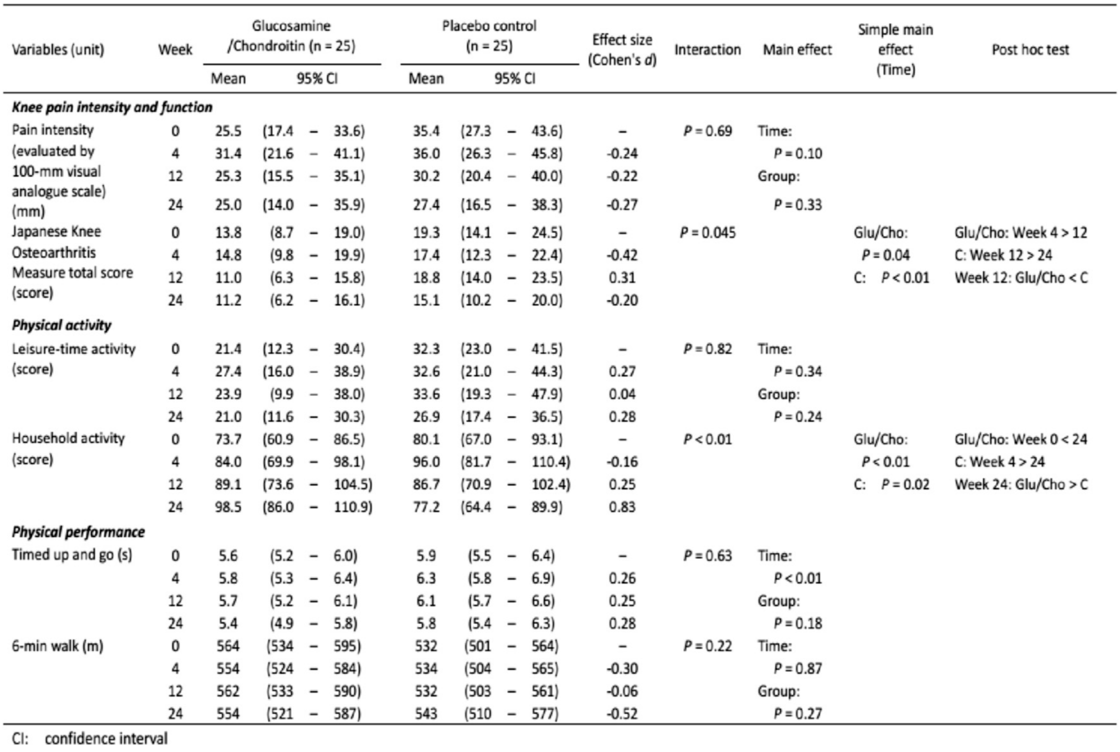


Figure 1. Percent change since baseline in the 2 primary outcomes in the glucosamine/chondroitin sulfate treatment group (blue) and the placebo group (red). (A) The placebo group had a greater decrease in pain than the treatment group, although this difference was not statistically significant. (B) The treatment and placebo group had near-identical decreases in osteoarthritis scores, although they took different paths to get there, which triggered a significant group x time effect.

shows a screenshot of Table 2 from their study [1], which presents the intention-to-treat results from the trial. The point of the table is to compare outcomes in the 2 randomization groups.

Had the table been well-focused, it would have been easier for readers—and the authors themselves—to notice that the groups had similar declines in pain and osteoarthritis score (Japanese Knee Osteoarthritis Measure) and that the final, 24-week effect sizes for both primary outcomes are negative: −0.27 for pain and −0.20 for osteoarthritis score. Negative effects indicate that the placebo group had a greater improvement than the treatment group.

Unfortunately, it is easy to miss this amidst the table's clutter. For example, the reader is bombarded with 48 different 95% confidence intervals, for the mean values of each outcome for each group at each time point. These add little to the reader's understanding of whether treatment beat placebo. The table also contains columns for main effects and simple main effects,



Variables (unit)	Week	Glucosamine /Chondroitin (n = 25)		Placebo control (n = 25)		Effect size (Cohen's *d*)	Interaction	Main effect	Simple main effect (Time)	Post hoc test
		Mean	95% CI	Mean	95% CI					
Knee pain intensity and function										
Pain intensity (evaluated by 100-mm visual analogue scale) (mm)	0	25.5	(17.4 – 33.6)	35.4	(27.3 – 43.6)	–	*P* = 0.69	Time:		
	4	31.4	(21.6 – 41.1)	36.0	(26.3 – 45.8)	-0.24		*P* = 0.10		
	12	25.3	(15.5 – 35.1)	30.2	(20.4 – 40.0)	-0.22		Group:		
	24	25.0	(14.0 – 35.9)	27.4	(16.5 – 38.3)	-0.27		*P* = 0.33		
Japanese Knee Osteoarthritis Measure total score (score)	0	13.8	(8.7 – 19.0)	19.3	(14.1 – 24.5)	–	*P* = 0.045		Glu/Cho:	Glu/Cho: Week 4 > 12
	4	14.8	(9.8 – 19.9)	17.4	(12.3 – 22.4)	-0.42			*P* = 0.04	C: Week 12 > 24
	12	11.0	(6.3 – 15.8)	18.8	(14.0 – 23.5)	0.31			C: *P* < 0.01	Week 12: Glu/Cho < C
	24	11.2	(6.2 – 16.1)	15.1	(10.2 – 20.0)	-0.20				
Physical activity										
Leisure-time activity (score)	0	21.4	(12.3 – 30.4)	32.3	(23.0 – 41.5)	–	*P* = 0.82	Time:		
	4	27.4	(16.0 – 38.9)	32.6	(21.0 – 44.3)	0.27		*P* = 0.34		
	12	23.9	(9.9 – 38.0)	33.6	(19.3 – 47.9)	0.04		Group:		
	24	21.0	(11.6 – 30.3)	26.9	(17.4 – 36.5)	0.28		*P* = 0.24		
Household activity (score)	0	73.7	(60.9 – 86.5)	80.1	(67.0 – 93.1)	–	*P* < 0.01		Glu/Cho:	Glu/Cho: Week 0 < 24
	4	84.0	(69.9 – 98.1)	96.0	(81.7 – 110.4)	-0.16			*P* < 0.01	C: Week 4 > 24
	12	89.1	(73.6 – 104.5)	86.7	(70.9 – 102.4)	0.25			C: *P* = 0.02	Week 24: Glu/Cho > C
	24	98.5	(86.0 – 110.9)	77.2	(64.4 – 89.9)	0.83				
Physical performance										
Timed up and go (s)	0	5.6	(5.2 – 6.0)	5.9	(5.5 – 6.4)	–	*P* = 0.63	Time:		
	4	5.8	(5.3 – 6.4)	6.3	(5.8 – 6.9)	0.26		*P* < 0.01		
	12	5.7	(5.2 – 6.1)	6.1	(5.7 – 6.6)	0.25		Group:		
	24	5.4	(4.9 – 5.8)	5.8	(5.4 – 6.3)	0.28		*P* = 0.18		
6-min walk (m)	0	564	(534 – 595)	532	(501 – 564)	–	*P* = 0.22	Time:		
	4	554	(524 – 584)	534	(504 – 565)	-0.30		*P* = 0.87		
	12	562	(533 – 590)	532	(503 – 561)	-0.06		Group:		
	24	554	(521 – 587)	543	(510 – 577)	-0.52		*P* = 0.27		

CI: confidence interval

Figure 2. Screenshot of Table 2 from the authors' study [1]. The table is cluttered with distracting and irrelevant details that detract from authors' and readers' ability to understand what the data show. With permission of Springer.

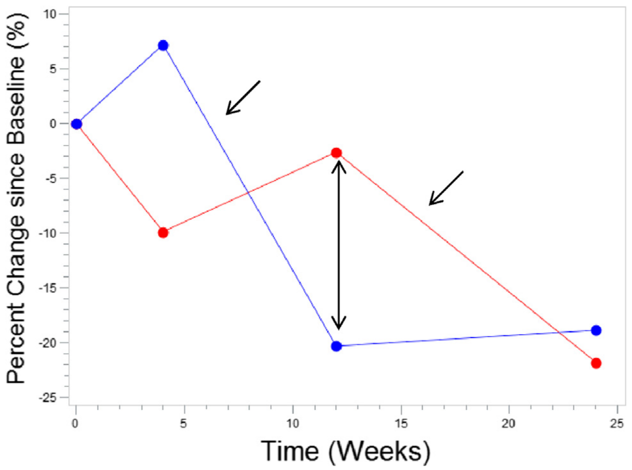


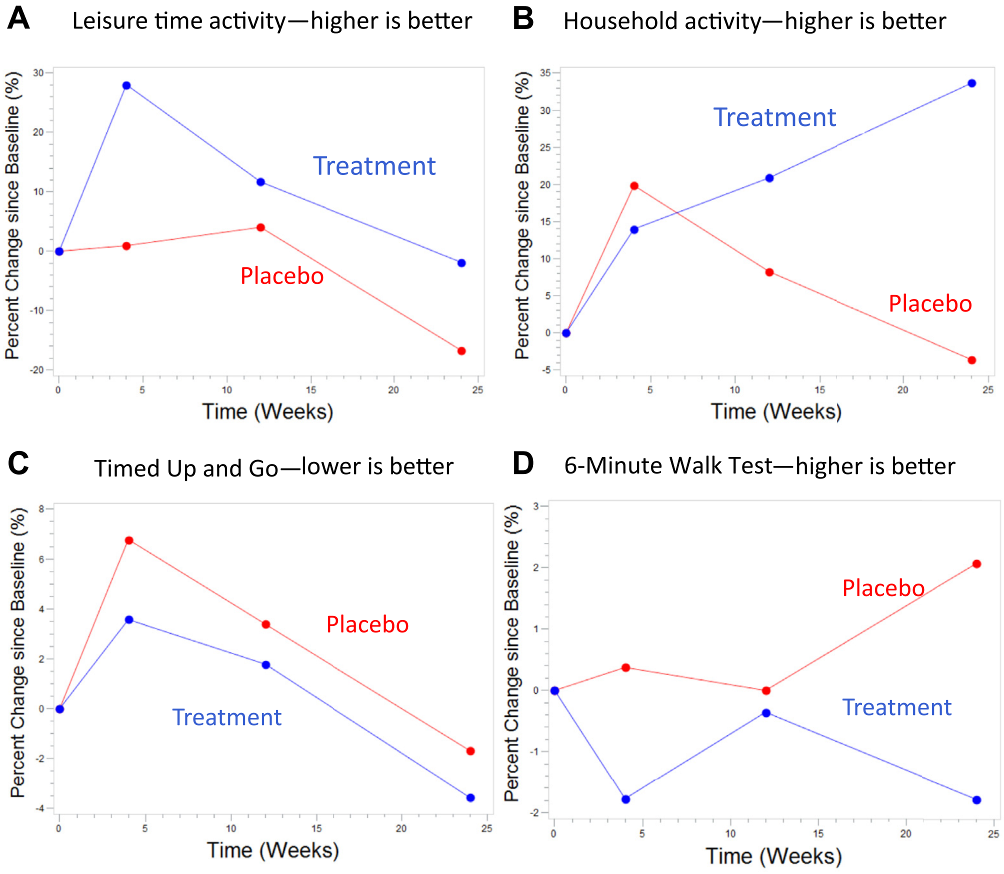
Figure 3. The authors ran 16 "post-hoc" tests for osteoarthritis levels—6 within-group comparisons for each group and 4 between-group comparisons at each time point. The 3 arrows indicate the 3 isolated "significant" declines or differences—there was a significant decrease from 4 to 12 weeks in the treatment group; a significant drop from 12 to 24 weeks in the placebo group; and a significant difference between treatment and placebo at 12 weeks. The overall declines in both groups may reflect a true placebo effect, but the between-group difference at 12 weeks is likely just a chance fluctuation since the reverse pattern is seen at 4 weeks and no difference is seen at 24 weeks.

which are irrelevant for comparing treatment with placebo. For example, a significant main effect for time just tells us that participants got better overall; and a significant main effect for time within one group just tells us that this group improved overall—not whether the 2 groups differed. It is too easy for readers and authors to get lost in all these numbers and miss the forest for the trees.

Consider the Totality of the Evidence

The authors also arrived at the wrong conclusion by cherry-picking and selectively reporting *P* values. When there are no effects, about 1 in 20 *P* values will come out significant at the .05 level just by chance [2]. Therefore, it is important to consider any significant *P* values in the context of how many statistical tests were run.

In this example, the authors calculated numerous *P* values for each outcome. For example, the "post-hoc" tests for osteoarthritis score included 16 tests: 6 within-group comparisons for each group—baseline versus 4 weeks, baseline versus 12 weeks, baseline versus 24



A Leisure time activity—higher is better

B Household activity—higher is better

C Timed Up and Go—lower is better

D 6-Minute Walk Test—higher is better

Figure 4. Percent change since baseline for the secondary outcomes by group. The groups did not differ significantly for leisure time activity (A), Timed Up and Go (C), or the 6-Minute Walk Test (D). The glucosamine/chondroitin sulfate treatment group did improve significantly more in household activity score compared with the placebo group (B). However, this could be a chance finding, given the large number of outcomes and time points considered. Of 6 total outcomes, one half favored the placebo by 24 weeks (pain, osteoarthritis score, 6-Minute Walk Test) and half favored the treatment group by 24 weeks (leisure activity, household activity, Timed Up and Go).

weeks, 4 weeks versus 12 weeks, 4 weeks versus 24 weeks, and 12 weeks versus 24 weeks—and 4 between-group comparisons at each time point.

Of these 16 comparisons, 3 came out significant (Figure 3). Two of these significant differences likely reflect a true placebo effect—both groups appear to decline overall in osteoarthritis severity, and this decline achieved statistical significance within one time interval per group (4-12 weeks for treatment, 12-24 weeks for placebo). However, the between-group difference at 12 weeks appears to be a chance fluctuation—after all, the reverse pattern is seen at 4 weeks and no difference is seen at 24 weeks (Figure 3).

In their abstract, the authors selectively report the 2 significant *P* values that involve the treatment group: "According to the post hoc test, it [osteoarthritis score] significantly decreased (ie, improved knee function) from the 4- to 12-week follow-up in the Glu/Cho group and the Glu/Cho group score was significantly lower than the C group at the 12-week follow-up." However, they ignore the significant decrease in the placebo group and the fact that the placebo group actually had a slightly greater decline in osteoarthritis score by 24 weeks.

Turning to the secondary outcomes, the authors did find one significant difference that favored the treatment group—the treatment group experienced a larger improvement in household activity score (Figure 4). This could reflect a real benefit, but it's also entirely consistent with a chance finding given that was just a secondary outcome and was 1 of 6 outcomes considered. In fact, I would say it is highly probable that this was a chance fluctuation—and I will illustrate why in the next section, where I consider the totality of the evidence in terms of effect sizes.

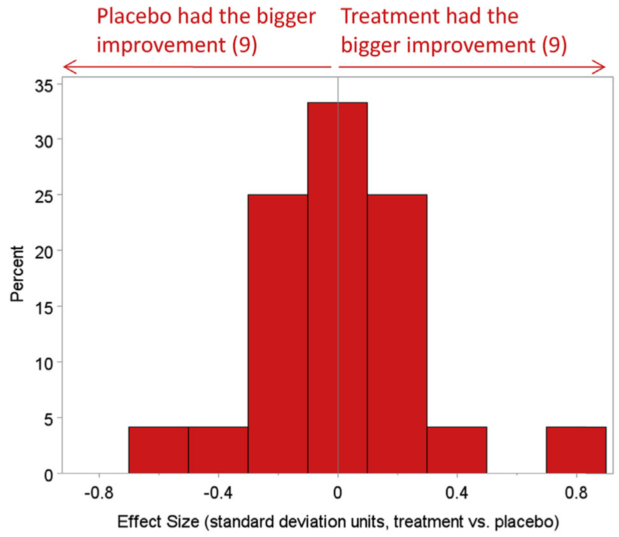


Figure 5. Histogram of the 18 effect sizes reported in Table 2 of the authors' study [1]. The effect size gives the difference in improvement in the treatment group versus the placebo group in standard deviation units. For example, an effect size of 0.50 means that treatment group improved by half a standard deviation more than the treatment group, whereas an effect size of −0.50 means the placebo group improved by half a standard deviation more than the treatment group. An effect size was calculated for each outcome and follow-up time point (total of 18). In 9 cases, the effect size favored the placebo group (<0); and in 9 cases the effect size favored the treatment group (>0).

Focus on Effect Sizes More Than *P* Values

In their discussion, the authors focus exclusively on *P* values and ignore effect sizes. Had they focused more on the magnitude and direction of effect sizes, they might have noticed something important—the distribution of effect sizes is exactly what we would expect to see if the treatment and placebo pills are equally effective.

In Table 2 of the study [1], the authors report 18 effect sizes—one for each outcome and each follow-up time point. The effect sizes are the difference in improvement in the treatment group minus the placebo group expressed in standard deviation units. For example, an effect size of 0.50 means that treatment group improved by a half standard deviation more than the treatment group, whereas an effect size of −0.50 means the placebo group improved by a half standard deviation more than the treatment group.

If you carefully scan the effect sizes in this table (Figure 2 here), you will notice that exactly half are negative and half are positive. Also, the magnitudes of the effect sizes in each direction are similar—with just 3 exceeding a third of a standard deviation difference. The placebo group has 2 effect sizes this big: −0.42 for osteoarthritis score at 4 weeks and −0.52 for the 6-Minute Walk Test at 24 weeks, whereas the treatment group has one: 0.83 for household activity at 24 weeks.

I plotted these 18 effect sizes in a histogram to make it easier to examine the evidence (Figure 5). If the null hypothesis is true (treatment is no better than placebo), we should see a symmetric, bell-shaped distribution around 0. Indeed, that is exactly what we see. One half of the effect sizes are below 0 and half are above 0; and the magnitudes in each direction are nearly identical. Although the largest effect size does belong to the treatment group, this seems likely just a chance fluctuation when viewed in the context of the other 17 effect sizes.

Conclusions

Science loses credibility when authors draw the wrong conclusion from their data. Here, I have presented an example in which the authors conclude that a supplement—glucosamine/chondroitin sulfate—is beneficial when, in fact, their data show a clear null result. In this case, the null finding is clinically informative—patients and doctors need to know when an intervention does not work so that people can avoid useless and expensive treatments. However, in this case, I suspect the authors' hopes and expectations for a positive result kept them from seeing the truth in their data. I have outlined 4 key steps that authors can take to avoid falling into this trap. Readers can also apply these 4 principles to help evaluate the evidence in published papers; they should pay close attention to any plots and effect sizes given, ignore distracting data in tables, and look at all the evidence rather than the isolated tidbits that authors choose to highlight. Note that these principles are fairly easy to implement and do not require extensive statistical training.

References

1. Tsuji T, Yoon J, Kitano N, Okura T, Tanaka K. Effects of N-acetyl glucosamine and chondroitin sulfate supplementation on knee pain and self-reported knee function in middle-aged and older Japanese adults: a randomized, double-blind, placebo-controlled trial. Aging Clin Exp Res 2016;28:197-205.
2. Sainani KL. The problem of multiple testing. PM R 2009;2: 1098-1103.

Disclosure

K.L.S. Department of Health Research and Policy, Stanford University, Stanford, CA.
Address correspondence to: K.L.S; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose