Clinical Versus Statistical Significance

Kristin L. Sainani, PhD

Many researchers focus so much on *P* values and statistical significance that they overlook a more important piece of information: effect size. Just because the *P* value is small and the result is highly statistically significant doesn't guarantee that the effect size is large. Big sample sizes give high resolution—so high that even minute differences between groups can be detected. These differences aren't a fluke; they are real. They are just so small that they aren't likely to have any meaningful impact. This article gives readers tips for spotting results that are statistically significant but clinically irrelevant.

CASE STUDY 1: EXERCISE AND WEIGHT GAIN

A 2010 study in the *Journal of the American Medical Association* concluded that women need 60 minutes of moderate exercise a day to prevent weight gain in middle age [1]. The study garnered widespread attention in the media, with headlines such as "New Exercise Goal: 60 Minutes a Day" (*Wall Street Journal*) and already has been cited in the medical literature more than 50 times. The study had many strengths: it was large (34,079 women), prospective, and lengthy (13 years of follow-up); it also used elegant regression models that incorporated a woman's changing exercise habits. I would argue, however, that the authors misinterpreted the results of this elegant analysis.

The researchers compared the average weight gain in 3 exercise groups: low ($<$7.5 metabolic equivalent [MET] hours/week), medium (7.5 to $<$21 MET hours/week), and high exercisers ($\geq$21 MET hours/week). They found that high exercisers gained significantly less weight than medium exercisers ($P = .003$) and low exercisers ($P = .002$). These *P* values are impressive, but the magnitude of the differences in weight gain might surprise you.

The results are given in 3-year intervals because a woman's exercise status was measured every 3 years and used to predict her weight change during the subsequent 3 years. During any 3-year period, high exercisers gained an average of just 0.12 kg—0.26 lb—less than low exercisers (and 0.11 kg less than medium exercisers, Table 1). Even projected over 13 years (the time frame of the study), high exercisers saved themselves an average of just 1 lb of weight gain. Does this payoff motivate you to put on your running shoes?

Besides regression modeling, the authors graphed the weight gain over time according to baseline exercise group (Figure 1). What's striking about this graphic is that the 3 lines are almost perfectly parallel; there is no difference in the pattern of weight gain over time in the 3 groups. High exercisers are much lighter than the other women at baseline, but they do not appear protected from weight gain longitudinally.

The visual is powerful and easy to interpret, and the authors correctly conclude: "Figure 2 shows the trajectory of weight gain over time by baseline physical activity levels. When classified by this single measure of physical activity, all 3 groups showed similar weight gain patterns over time." However, the authors mistakenly assume that the visual would have looked different had their graph encapsulated the flux in exercise groups. In fact, I do not believe this would be the case. A difference in the slopes of the lines of just 0.11 or 0.12 kg per 3 years will barely be perceptible. Just consider the first 3 years of the study. The physical activity at baseline should predict the weight-gain trajectory during the first 3 years. But do those lines (0 to 36 months) look like they have different slopes to you?

CASE STUDY 2: EXERCISE AND DRINKING

A 2009 study in the *American Journal of Health Promotion* reported an association between drinking alcohol and exercising on the basis of national survey data [2]. This study was a

K.L.S. Department of Health Research and Policy, Division of Epidemiology, Stanford University, HRP Redwood Building, Stanford, CA 94305. Address correspondence to: K.L.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 18, 2012; accepted April 18, 2012.

PM&R
1934-1482/12/$36.00
Printed in U.S.A.

Vol. 4, 442–445, June 2012
http://dx.doi.org/10.1016/j.pmrj.2012.04.014

Table 1. *The average difference in weight gain (in kilograms) over any 3-year period in the medium- and low-exercise groups compared with the high-exercise group (reference group)*

Group	No. of Women†	Physical Activity, MET Hours per Week*			*P* Value for Trend	*P* Value for Interaction
		<7.5	7.5 to < 21	≥21		
All women						
Analytical model‡						
1		0.15 (0.04)	0.12 (0.04)	0 (Reference)	<.001	
2		0.12 (0.04)	0.11 (0.04)	0 (Reference)	<.001	
Age, y						
<55	21,363	0.12 (0.08)	0.02 (0.08)	0 (Reference)	<.001	
55-64	9699	0.24 (0.06)	0.19 (0.06)	0 (Reference)	<.001	<.001
≥65	3017	−0.09 (0.07)	0.07 (0.07)	0 (Reference)	.13	
BMI						
<25.0	17,475	0.21 (0.04)	0.14 (0.04)	0 (Reference)	<.001	
25-29.9	10,516	−0.04 (0.06)	−0.04 (0.06)	0 (Reference)	.56	<.001
≥30.0	6088	0.16 (0.14)	0.13 (0.16)	0 (Reference)	.50	
Smoking status						
Never	17,692	0.18 (0.05)	0.17 (0.05)	0 (Reference)	<.001	
Former	12,169	0.06 (0.06)	0.05 (0.06)	0 (Reference)	.04	.53
Current	4186	0.15 (0.15)	0.12 (0.16)	0 (Reference)	.11	
Menopausal status						
Premenopausal	9821	0.19 (0.13)	0.08 (0.13)	0 (Reference)	.03	.04
Postmenopausal	17,762	0.12 (0.04)	0.12 (0.04)	0 (Reference)	<.001	

MET = metabolic equivalent; BMI = body mass index, which is calculated as weight in kilograms divided by height in meters squared.

Reproduced with permission from Lee MI, Djoussé L, Sesso HD, Wang L, Buring JE. Physical activity and weight gain prevention. JAMA 2010;303:1173-1179.

*The mean (SD) difference in weight in kilograms is compared with the reference group. The mean (SD) interval during which weight change was assessed was 2.88 (0.11) years. An expenditure of 7.5 MET hours per week is equivalent to 150 minutes per week of moderate-intensity physical activity, the minimum recommended by the federal government (1); 21 MET hours per week is equivalent to 60 minutes per day (420 min/wk) of moderate-intensity physical activity, recommended by the Institute of Medicine (1).

†Those in the group at baseline.

‡Model 1 was adjusted for age, baseline weight, height, and time interval between weight assessments. Model 2 was additionally adjusted for race, educational attainment, smoking status, menopausal status, hormone-replacement therapy use, hypertension, diabetes, alcohol consumption, and quintiles of intakes of total energy, saturated fat, and fruits and vegetables. Analyses according to subgroups of women all used estimates from model 2.

cross-sectional study, and thus the authors did not claim to prove causation (or to know the direction of causation), but the findings still were covered widely in the news media. The study concluded that "our results strongly suggest that alcohol consumption and physical activity are positively correlated."

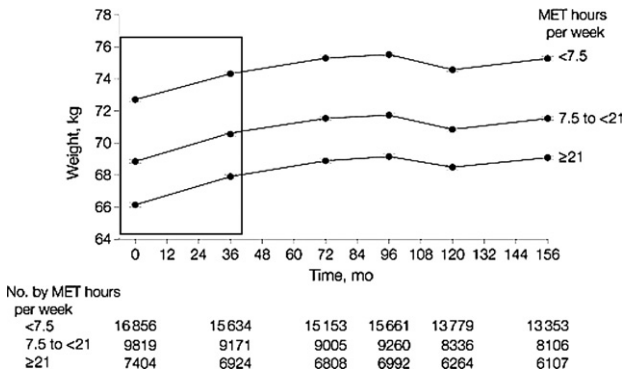

Figure 1. The trajectory of weight gain over time by the *baseline* exercise group (high, medium, and low). The box has been added to highlight the effect in the first 3 years. *MET* = metabolic equivalent. Reproduced with permission from Lee MI, Djoussé L, Sesso HD, Wang L, Buring JE. Physical activity and weight gain prevention. JAMA 2010;303:1173-1179.

Indeed, the observed association between alcohol and exercise was statistically significant ($P < .01$). But, again, the magnitude of the association may surprise you. For women, current drinkers exercised a whopping 7.2 more minutes more *per week* than abstainers. Also, drinking 10 extra drinks per months was associated with an extra 2.2 minutes per week of exercise (9.6 minutes per month!). In other words, 1 extra drink per month was linked to about 1 extra minute of exercise per month. As one of my students surmised, maybe they got the extra exercise walking to the fridge and back to grab the beer?

HUGE SAMPLES CHANGE THE GAME

So what's going on in the aforementioned examples? How can such small effects be statistically significant? The answer is that both studies have enormous sample sizes. In the first study, 34,079 women each contributed multiple measurements to the analysis, for a total far exceeding 100,000 observations. In the second study, the survey included more than 230,000 people. Huge samples are fantastic, but they make *P* values obsolete. *P* values are a tool for separating real

Table 2. *The minimum correlation coefficient that will be statistically significant for various sample sizes**

Sample Size	Minimum Correlation Coefficient That Will Be Statistically Significant, $P < .05$
10	.63
100	.20
1000	.06
10,000	.02
100,000	.006
1,000,000	.002

*Calculated using the approximation $r = \frac{2}{\sqrt{n}}$.

effects from chance variation, but chance variation isn't an issue for huge samples—random fluctuation simply gets drowned out.

A significant *P* value ($P < .05$) tells us that we can rule out a null effect (eg, a difference of 0) with 95% certainty. However, large samples give such precision that ruling out the null value doesn't mean a whole lot. We may be able to rule out 0, but this does not guarantee that the effect is far from 0. The confidence interval (which is very narrow with huge samples) might range from 0.01-0.02, for example.

In fact, any effect size—no matter how small—can be made statistically significant if the sample size is large enough. Consider a Pearson correlation coefficient, which is the measure of the strength of the association between 2 variables. Whether a correlation coefficient achieves statistical significance depends on only 2 things: (1) the size of the correlation, from -1 to $+1$; and (2) the sample size. The minimum correlation coefficient that will achieve statistical significance ($P < .05$) for a given sample size is approximately (see the "In Depth" box for derivation):

$$r = \frac{2}{\sqrt{n}}$$

When we inspect this formula, it is easy to see that with gigantic sample sizes, tiny correlation coefficients will be statistically significant. Specific examples are provided in Table 2. With a sample size of 100,000, a correlation coefficient of 0.006 will be statistically significant, but that doesn't make it a meaningful association.

This discussion has a caveat. *P* values depend on the size of the smallest group analyzed rather than the overall sample size. For example, if a study has 100,000 participants but the outcome develops in only 100 people, then chance variation still may play a large role in the observed effects. Thus if the overall sample size is huge but some subgroups of interest are small or the outcome is rare, then *P* values still may be relevant.

HOW DO YOU JUDGE CLINICAL SIGNIFICANCE?

One of the reasons that readers and authors alike love *P* values is that they give a simple, objective yes/no answer. In contrast, the concept of "clinical significance" is squishy; different people may disagree as to what constitutes a clinically important effect. However, *P* values simply are not helpful with huge sample sizes (as the aforementioned examples demonstrate); thus, like it or not, one has to focus on clinical significance.

To judge clinical significance, one needs to consider the 95% confidence interval. Statistical significance asks whether the confidence interval excludes the null value. In contrast, clinical significance asks whether any of the values in the confidence interval are big enough to care about. In example 1, the 95% confidence interval for the difference in weight gain between low exercisers and high exercisers was 0.04 to 0.20 kg (0.09 to 0.44 lb). This 95% confidence interval means that the effect plausibly could be as large as a 0.44-lb reduction in weight gain over 3 years.

Is this reduction large enough to care about? I would guess that most people would say no. In any case, the conclusion that "women should exercise an hour a day to prevent middle-age weight gain" is clearly overblown. Exercising an hour a day undoubtedly has many health benefits, but this study suggests that it has only minimal impact on weight-gain prevention in middle-aged women. (Of course, we can't rule

In Depth: The minimal correlation coefficient that will be significant for a given sample size.
The statistical test that is used to determine the P value for a correlation coefficient is:

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

Where r = correlation coefficient; n = sample size; and T_{n-2} = the *T* statistic (bigger *T* values correspond to smaller *P* values).
This formula can be algebraically rearranged to isolate r:

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

T values of approximately 2 correspond to $P = .05$ (which can be found by consulting a *T*-distribution table); so we can plug in 2 to get the approximate value of r that will be statistically significant for a given sample size:

$$r = \frac{2}{\sqrt{n-2+4}} = \frac{2}{\sqrt{n-2}} \approx \frac{2}{\sqrt{n}}$$

out the possibility that biases in the study masked a larger effect.)

For the study on exercise and drinking, the 95% confidence interval for the difference in weekly exercise between the drinkers and nondrinkers was 4.9 to 9.5 minutes, so the effect plausibly may be as large as 9.5 minutes more exercise per week. Is that number large enough to care about? Again, I would guess that most of us would say no.

CONCLUSIONS

With the availability of electronic medical records and other sources of large-scale data, huge studies are becoming more and more common. Thus the distinction between clinically significant and statistically significant results is going to have increasing relevance. Readers should be wary when studies involving tens of thousands or hundreds of thousands of participants boast impressive *P* values (unless the studies involve rare outcomes or exposures). The pertinent question for these studies is: Are any of the values within the 95% confidence interval big enough to care about? If the answer is no, then the effect is clinically insignificant and statistical significance is immaterial.

REFERENCES

1. Lee MI, Djoussé L, Sesso HD, Wang L, Buring JE. Physical activity and weight gain prevention. JAMA 2010;303:1173-1179.
2. French MT, Popovici I, Maclean JC. Do alcohol consumers exercise more? Findings from a national survey. Am J Health Promot 2009; 24:2-10.