“Top 10” statistical errors

JANUARY 25, 2020

KRISTIN SAINANI, PHD
KCOBB@STANFORD.EDU, @KRISTINSAINANI

“Top ten” statistical errors

1. Simple math and data errors
2. Study design mismatched to statistical question
3. Chance findings
4. Clinically irrelevant effect sizes
5. Exaggerated effect sizes
6. Wrong comparisons
7. Failure to account for correlated observations
8. Misinterpretations of null effects
9. Residual/unmeasured confounding
10. Spurious correlations/overfitting

1. Simple math and data errors

Simple math errors are surprisingly common!

2+2 = 5

Examples:

-N’s don’t add up

-Simple calculation errors

-Descriptive statistics don’t make sense

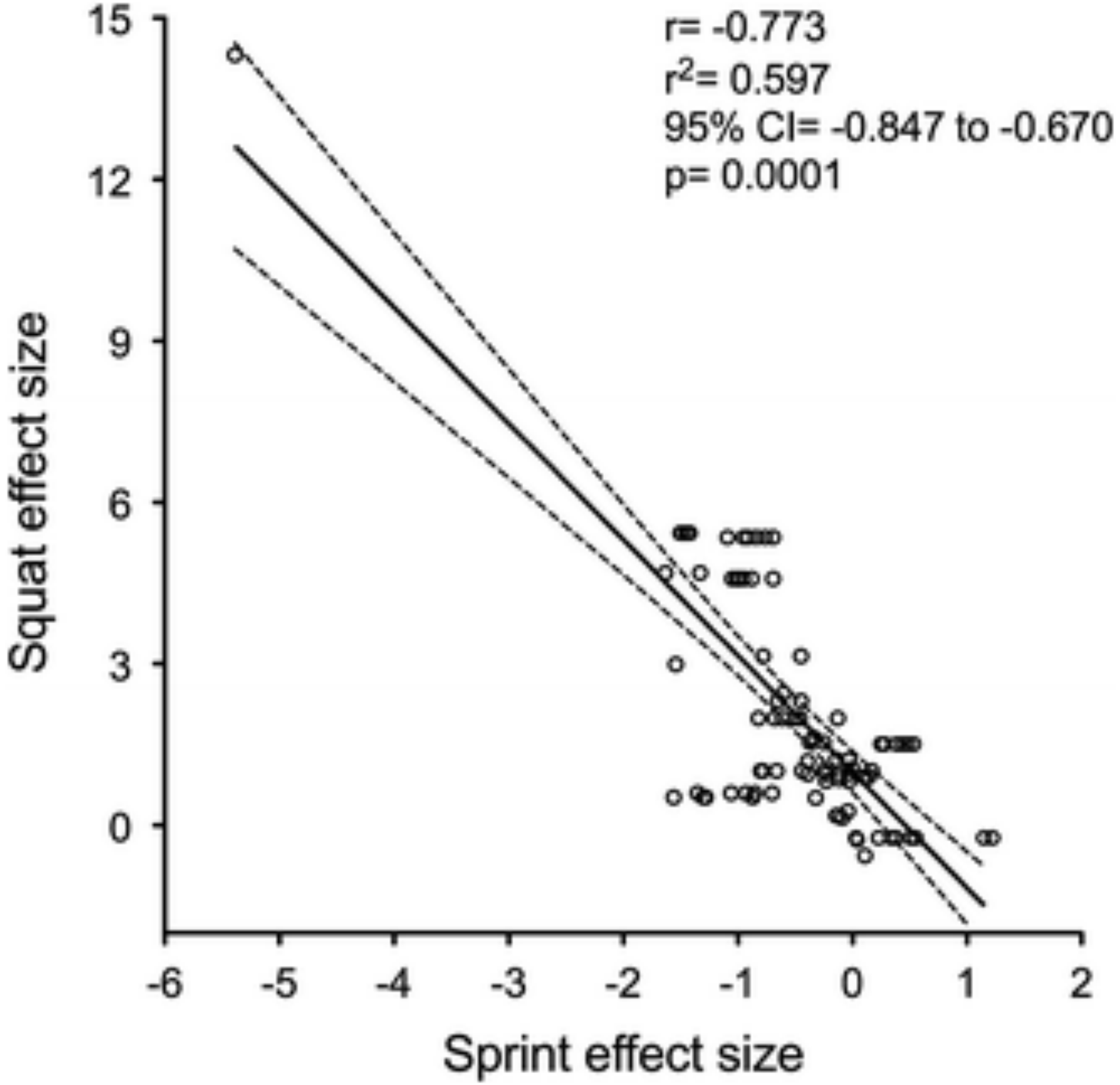
-Data in different tables or figures don’t match

-Other data errors that may be more hidden

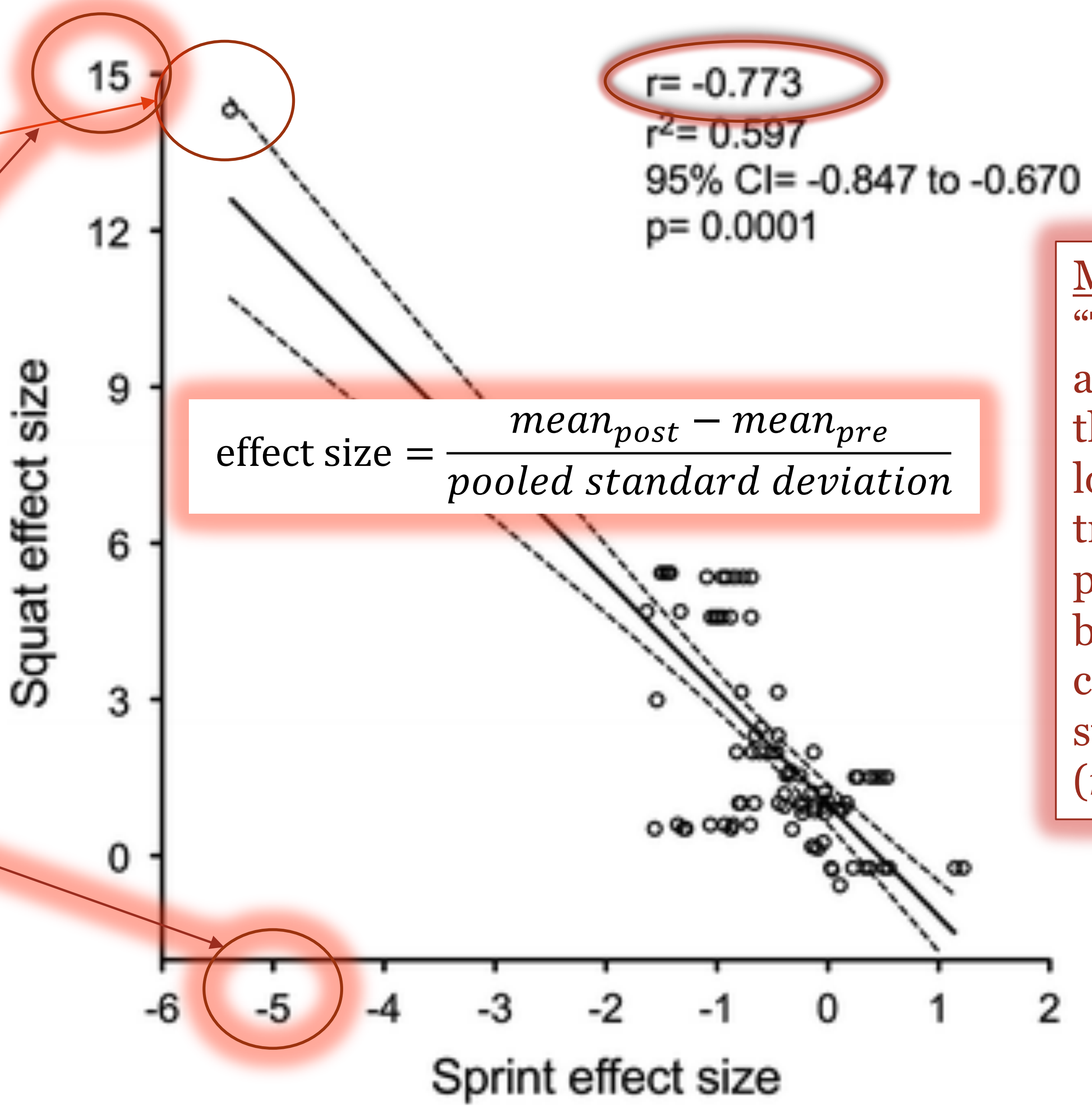
-Meta-analysis of studies of athletes who underwent a short training intervention (8 to 12 weeks)

-Athletes had leg strength (squatting ability) and sprint times measured before and after the intervention.

-Goal was to estimate correlation between improvement in leg strength and improvement in sprinting ability.



Seitz, L.B., Reyes, A., Tran, T.T. *et al.* Increases in Lower-Body Strength Transfer Positively to Sprint Performance: A Systematic Review with Meta-Analysis. *Sports Med* **44,** 1693–1702 (2014).



“Rogue data point”

Effect size of ~15 standard deviations!

Effect size of ~5 standard deviations!

$$\text{effect size} = \frac{mean_{post} - mean_{pre}}{pooled\ standard\ deviation}$$

Main study conclusion:
“The present meta-analysis suggests that there is a transfer of lower-body strength training to sprint performance as indicated by the very large correlation between squat strength ES and sprint ES ($r = -0.77$; $p \leq 0.001$).”

Seitz, L.B., Reyes, A., Tran, T.T. *et al.* Increases in Lower-Body Strength Transfer Positively to Sprint Performance: A Systematic Review with Meta-Analysis. *Sports Med* **44,** 1693–1702 (2014).

How big is a 15-standard deviation improvement in squatting ability?

A. a small effect

B. a medium effect

C. a large effect

D. a very large effect

E. an implausibly large effect

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

How big is a 15-standard deviation improvement?

For squats, baseline mean and standard deviation were 123 kg +/- 8 kg.

So, a 15-standard deviation improvement would mean that o*n averag*e athletes increased their squat ability from 123 kg to 243 kg.

What went wrong?

The authors extracted means and *STANDARD ERRORS* from the original paper. They plugged *STANDARD ERRORS* rather than *STANDARD DEVIATIONS* into the effect size formula.

Pre-intervention mean and SEM = 123 kg +/- 1.9 kg
Post-intervention mean and SEM = 148 kg +/- 1.5 kg → +25 kg / 1.7 kg = 14.7

Pre-intervention mean and SD = 123 kg +/- 8.5 kg
Post-intervention mean and SD = 148 kg +/- 6.7 kg → +25 kg / 7.6 kg = 3.3

2. Study design mismatched to statistical question

Researchers failed to choose the correct study design for the statistical question of interest, resulting in a study that is unable to answer the question of interest.

Examples:

-power calculation done for the wrong statistical test, resulting in the study being underpowered

-study lacks an appropriate control group to answer the question of interest

-superiority study was run when equivalence or non-inferiority was of interest

Example

<u>Aim:</u> The purpose of this study was to identify the clinical effectiveness of **oral versus topical NSAIDs** in the treatment of greater trochanteric pain syndrome.

<u>Methods:</u> A retrospective chart review of **25 patients** diagnosed with greater trochanteric pain syndrome were categorized into two groups: those who received **oral etodolac 400 milligrams** twice daily versus **topical diclofenac 3%** one gram tw

<u>Outcome</u>: Pain scores using the nu e, two-week, and six-week follow-up visits.

But lack of a statistically significant difference is not proof of non-inferiority!

<u>Results:</u> At **two weeks**, there was a statistically significant improvement in pain in both the oral and topical NSAID groups, **with no statistically significant difference between the groups (p=0.77).**

Similarly, at **six weeks**, there was a statistically significant improvement in pain in both the oral and topical NSAID groups, **with no statistically significant difference between the groups (p=0.59).**

<u>Conclusion:</u> Based on this study, the use of topical NSAIDs is non-inferior to oral NSAIDs in the treatment of GTPS.

J Pain, 16 (2015), p. s67

Study design was mismatched to statistical question…

The goal of the study was to show that topical NSAIDs are “no worse than” oral NSAIDs in the treatment of this pain syndrome. In other words, the goal was to show non-inferiority.

But the study was not designed as a non-inferiority study.

In a non-inferiority study, one must pre-specify a margin of equivalence and calculate sample size needs based on a non-inferiority design.

3. Chance findings

Are the authors cherry-picking results or engaging in p-hacking?

Examples:

-running many statistical tests (many endpoints or time points) but only highlighting the few that come out significant

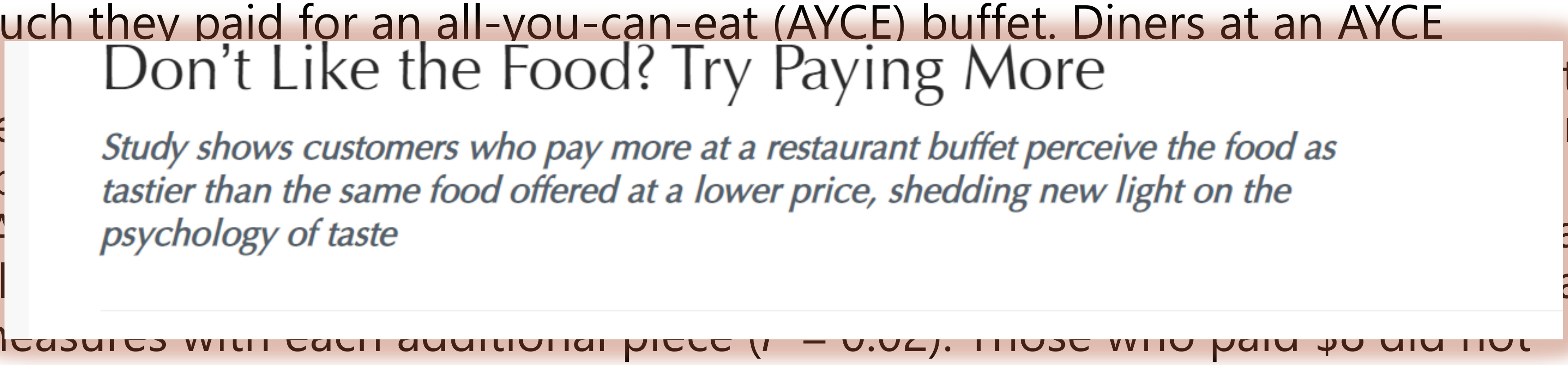
-intentionally or unintentionally manipulating data to get p<.05

s

Example

Abstract

A field experiment was conducted to assess how diners' taste evaluations change based on how much they paid for an all-you-can-eat (AYCE) buffet. Diners at an AYCE restaurant tion of each pie nd self-percep variance (A a as less tasty, l ach of these measures with each additional piece (P = 0.02). Those who paid $8 did not experience the same decrement in taste, satisfaction and enjoyment. Paying less for an AYCE experience may face the unintended consequence of food that is both less enjoyable and rapidly declining in taste and enjoyability. In a sense, AYCE customers get what they pay for.



Don’t Like the Food? Try Paying More

Study shows customers who pay more at a restaurant buffet perceive the food as tastier than the same food offered at a lower price, shedding new light on the psychology of taste

Lower buffet prices lead to less taste satisfaction. Journal of sensory studies. 2014 Oct;29(5):362-70.

Example

Methods

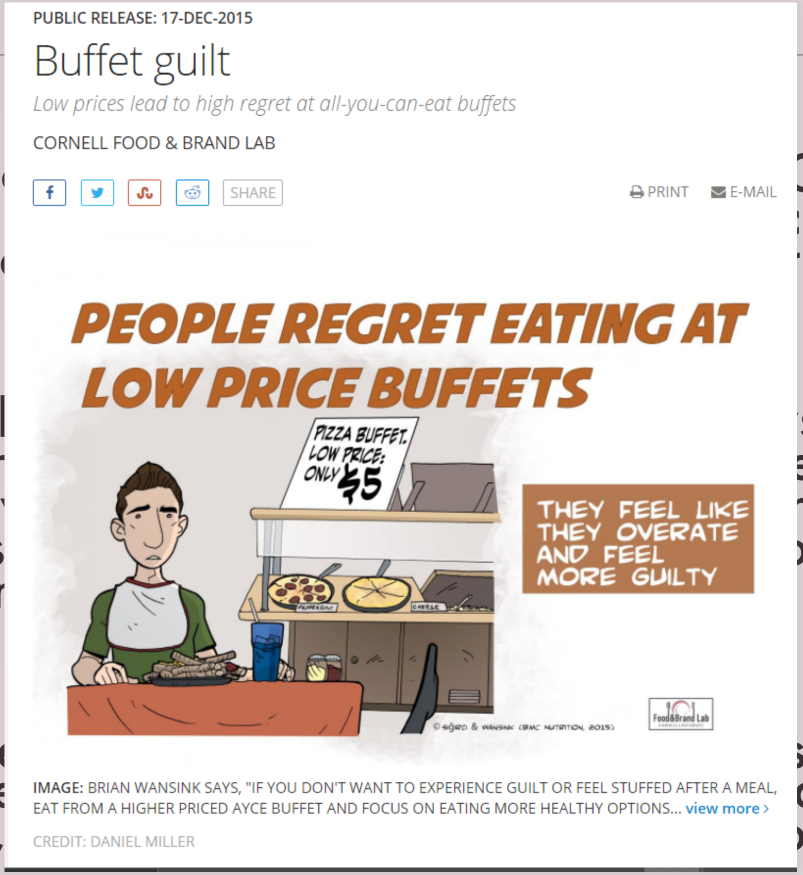
A field experiment was c
CE restaurant were either
charged $4 or $8 for an
h, participants rated
dimensions such as physi
they overate, and guilt.

Results

Diners who paid $4 for t
sically more uncomfortable
and had eaten more than
e diners who paid $8 for the
buffet ($p < 0.05$). Diners
higher ratings to overeating,
feelings of guilt and phys
paid $8 for the buffet, even if
they ate the exact same r

Conclusion

Paying less for an AYCE
sing consequences; lower
paying diners feel themse
able and guiltier compared to
the higher paying diners,
ount.



PUBLIC RELEASE: 17-DEC-2015

Buffet guilt

Low prices lead to high regret at all-you-can-eat buffets

CORNELL FOOD & BRAND LAB

SHARE PRINT E-MAIL

PEOPLE REGRET EATING AT LOW PRICE BUFFETS
PIZZA BUFFET, LOW PRICE: ONLY $5
THEY FEEL LIKE THEY OVERATE AND FEEL MORE GUILTY
© Sığırcı & Wansink (BMC Nutrition, 2015)
Food&Brand Lab

IMAGE: BRIAN WANSINK SAYS, "IF YOU DON'T WANT TO EXPERIENCE GUILT OR FEEL STUFFED AFTER A MEAL, EAT FROM A HIGHER PRICED AYCE BUFFET AND FOCUS ON EATING MORE HEALTHY OPTIONS... view more ›

CREDIT: DANIEL MILLER

Siğirci, Ö., Wansink, B. *BMC Nutr* **1,** 36 (2015). https://doi.org/10.1186/s40795-015-0030-x

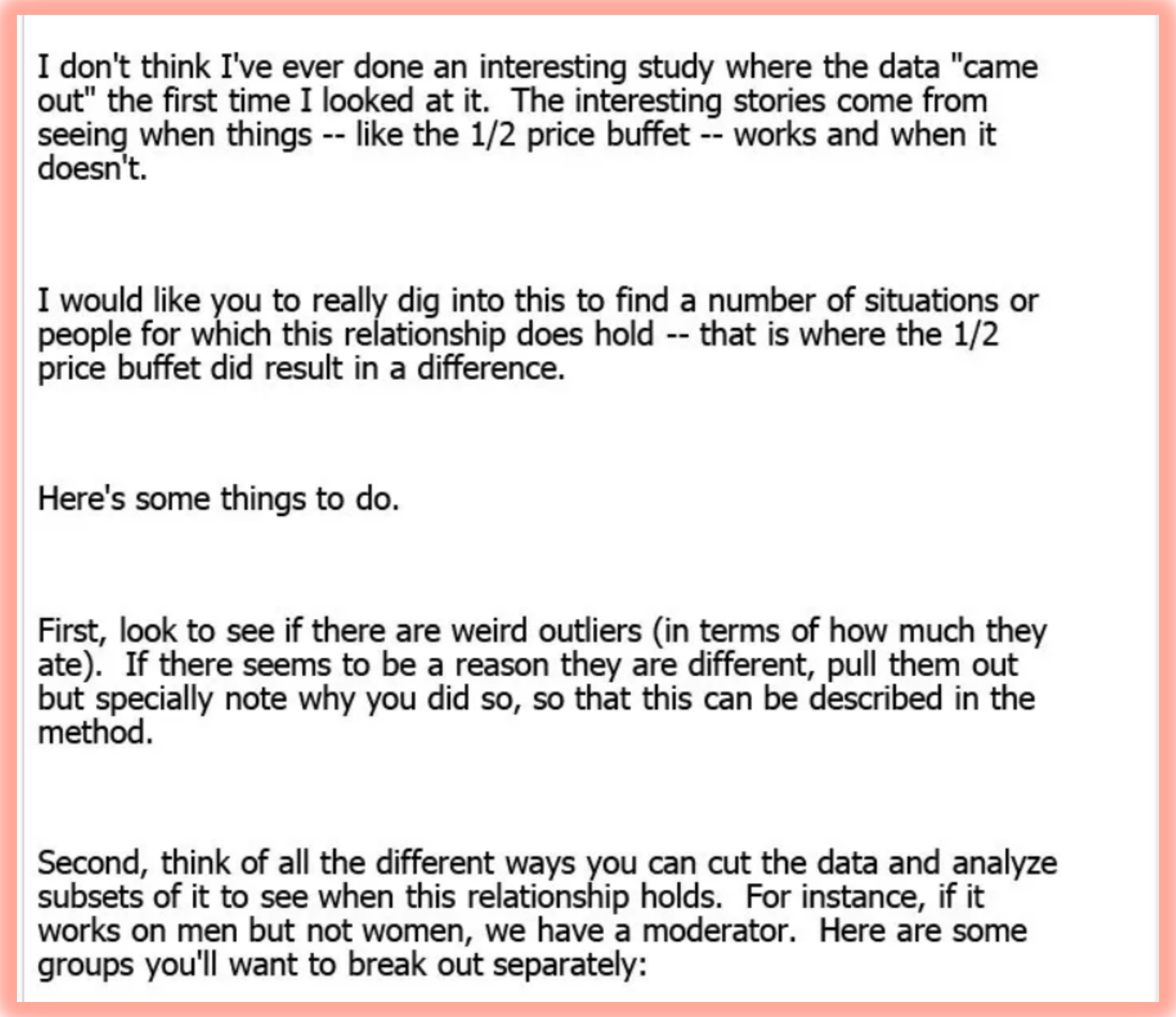
Blog post: “The grad Student who never said no.”

A PhD student from a Turkish university called to interview to be a visiting scholar for 6 months. Her dissertation was on a topic that was only indirectly related to our Lab's mission, but she really wanted to come and we had the room, so I said "Yes."

When she arrived, **I gave her a data set of a self-funded, failed study which had null results** (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people ½ as much as others). **I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed).** I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

Every day we would scratch our heads, ask "Why," **and come up with another way to reanalyze the data with yet another set of plausible hypotheses...**

Related email…



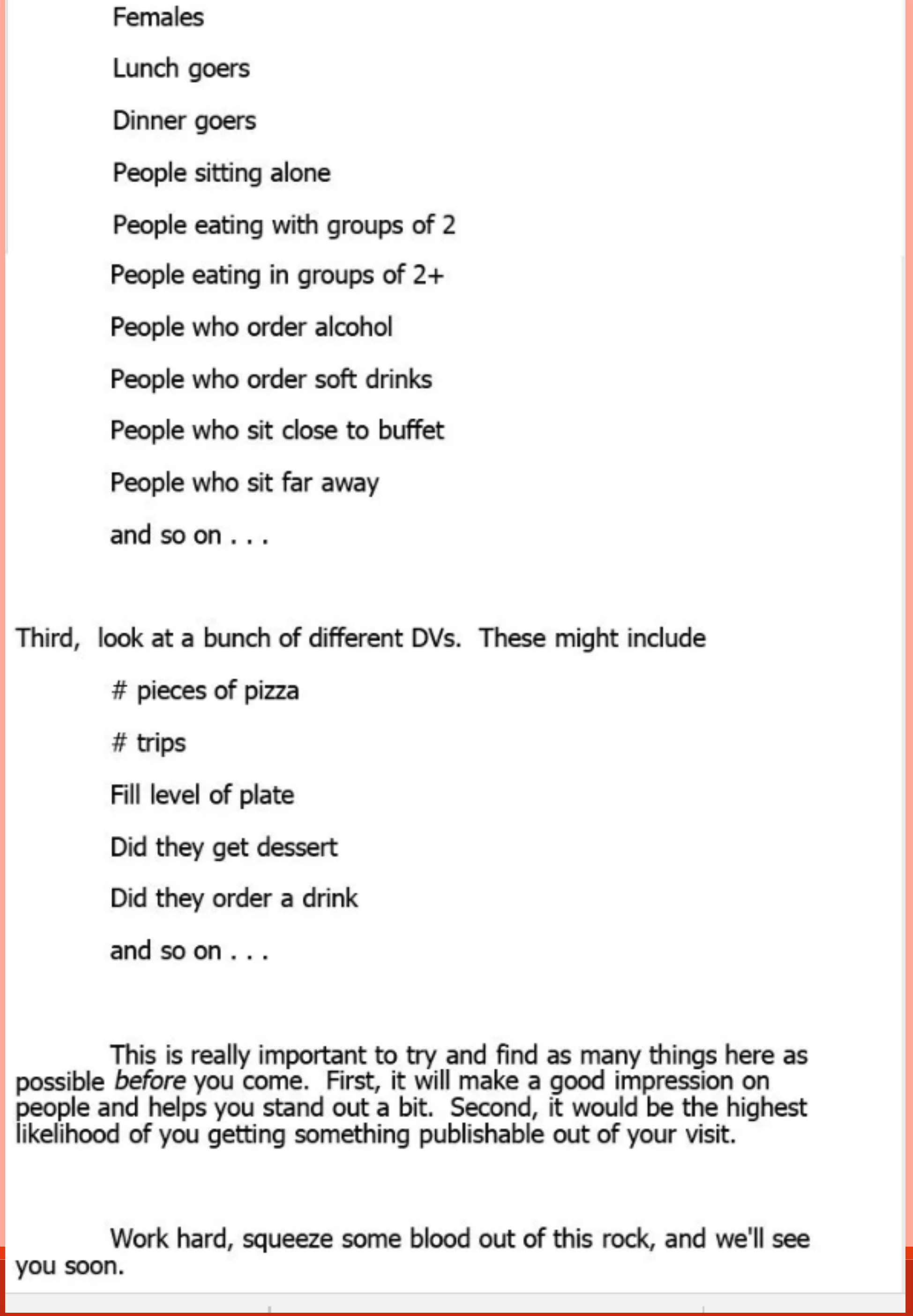
I don't think I've ever done an interesting study where the data "came out" the first time I looked at it. The interesting stories come from seeing when things -- like the 1/2 price buffet -- works and when it doesn't.

I would like you to really dig into this to find a number of situations or people for which this relationship does hold -- that is where the 1/2 price buffet did result in a difference.

Here's some things to do.

First, look to see if there are weird outliers (in terms of how much they ate). If there seems to be a reason they are different, pull them out but specially note why you did so, so that this can be described in the method.

Second, think of all the different ways you can cut the data and analyze subsets of it to see when this relationship holds. For instance, if it works on men but not women, we have a moderator. Here are some groups you'll want to break out separately:



Females

Lunch goers

Dinner goers

People sitting alone

People eating with groups of 2

People eating in groups of 2+

People who order alcohol

People who order soft drinks

People who sit close to buffet

People who sit far away

and so on . . .

Third, look at a bunch of different DVs. These might include

pieces of pizza

trips

Fill level of plate

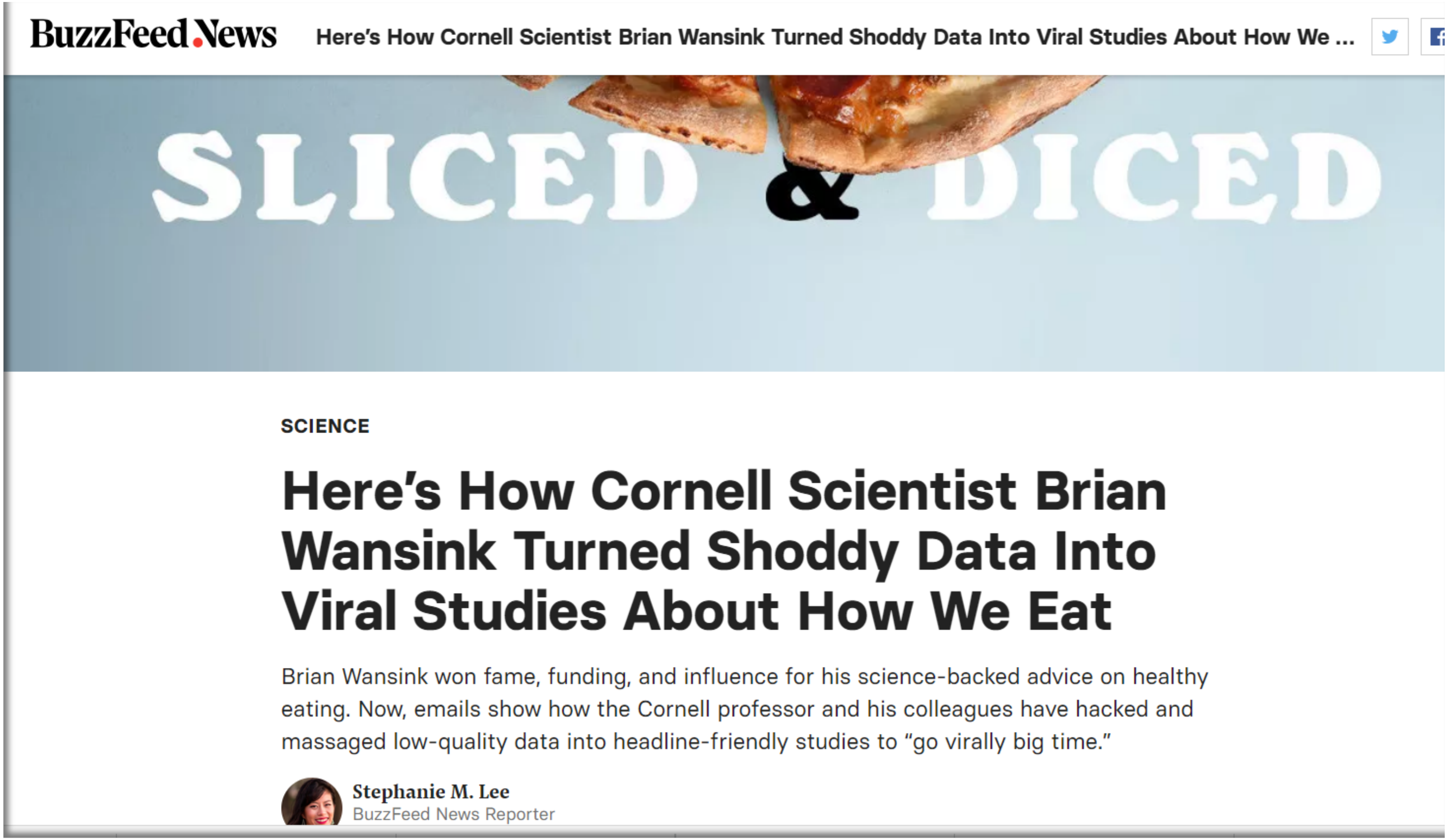
Did they get dessert

Did they order a drink

and so on . . .

This is really important to try and find as many things here as possible *before* you come. First, it will make a good impression on people and helps you stand out a bit. Second, it would be the highest likelihood of you getting something publishable out of your visit.

Work hard, squeeze some blood out of this rock, and we'll see you soon.



Here's How Cornell Scientist Brian Wansink Turned Shoddy Data Into Viral Studies About How We ...

SCIENCE

Here's How Cornell Scientist Brian Wansink Turned Shoddy Data Into Viral Studies About How We Eat

Brian Wansink won fame, funding, and influence for his science-backed advice on healthy eating. Now, emails show how the Cornell professor and his colleagues have hacked and massaged low-quality data into headline-friendly studies to "go virally big time."

Stephanie M. Lee
BuzzFeed News Reporter

4. Clinically irrelevant effect sizes

Trivial effects may achieve statistical significance if the sample size is large enou

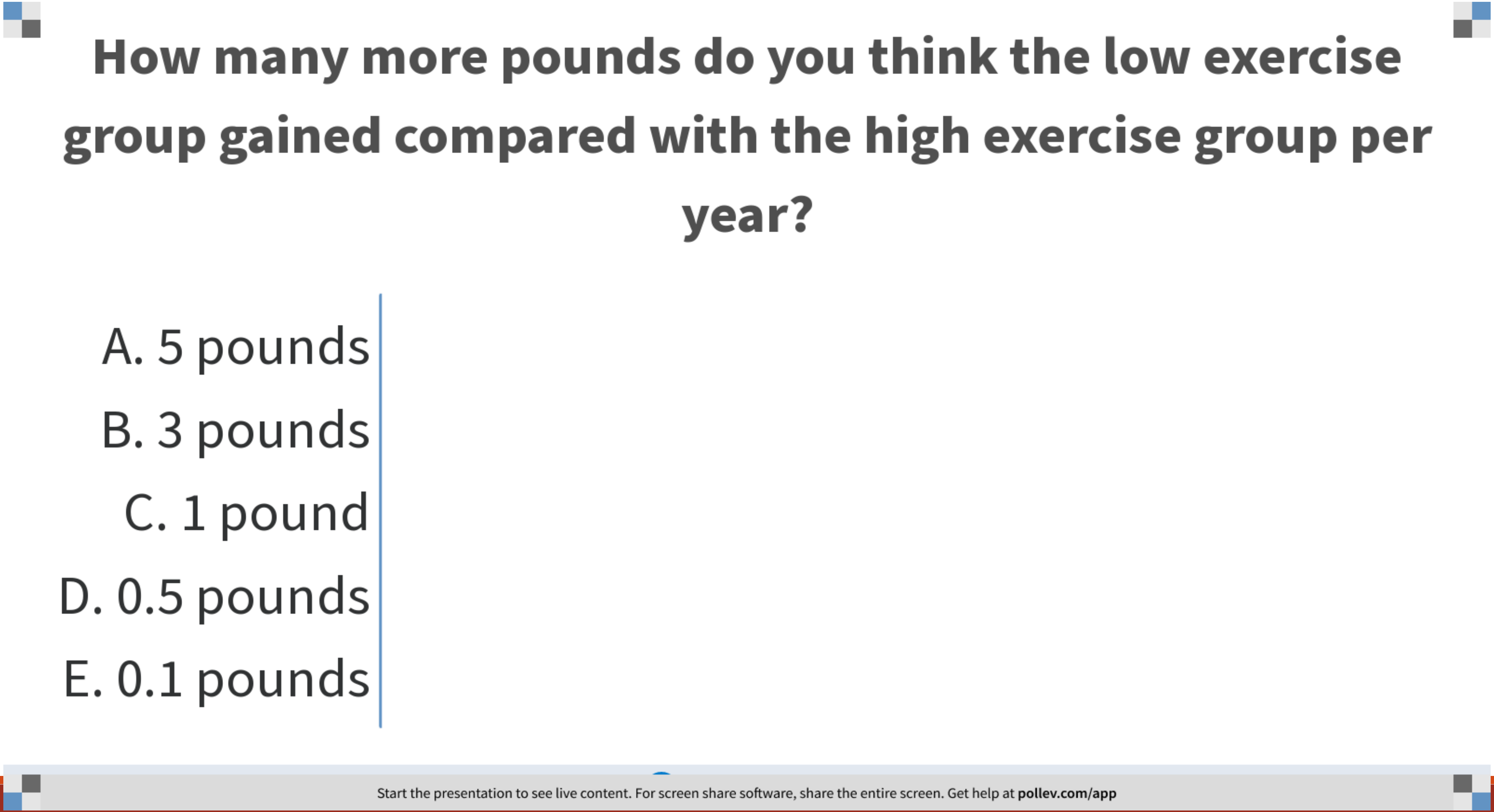
Example

A prospective cohort study of 34,079 women found that women who exercised >21 MET hours per week (≈60 minutes moderate-intensity exercise daily) gained **significantly less** weight than women who exercised <7.5 MET hours **(p<.001)**

Widely covered in the media. Headlines:

- "To Stay Trim, Women Need an Hour of Exercise Daily."
- "New Exercise Goal: 60 Minutes a Day"

Physical Activity and Weight Gain Prevention. JAMA 2010;303:1173-1179.



How many more pounds do you think the low exercise group gained compared with the high exercise group per year?

A. 5 pounds

B. 3 pounds

C. 1 pound

D. 0.5 pounds

E. 0.1 pounds

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992-2007a

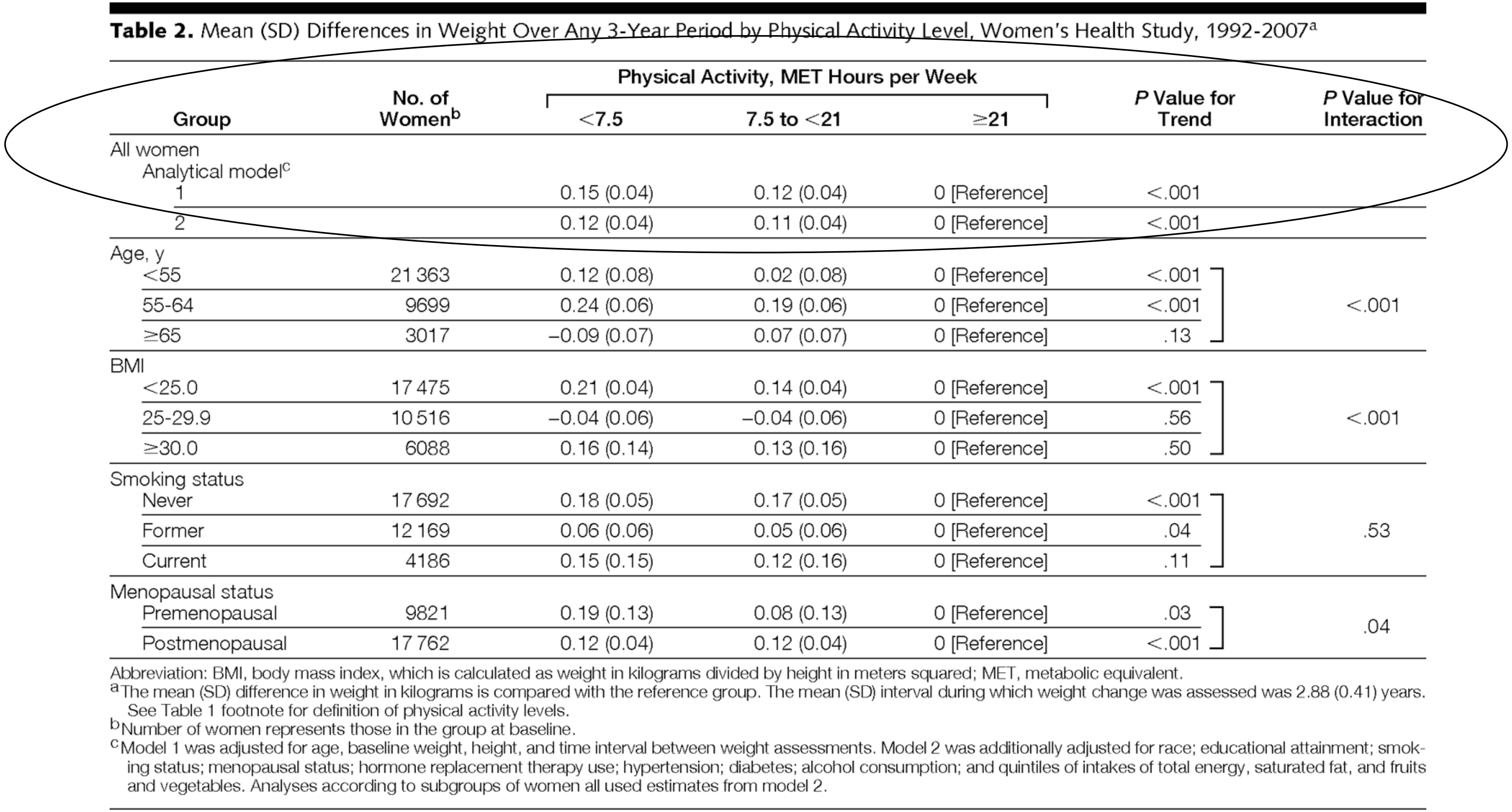


Table 2. Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992-2007[a]

Group	No. of Women[b]	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[c]						
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	<.001
55-64	9699	0.24 (0.06)	0.19 (0.06)	0 [Reference]	<.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	<.001
25-29.9	10 516	−0.04 (0.06)	−0.04 (0.06)	0 [Reference]	.56	
≥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	.53
Former	12 169	0.06 (0.06)	0.05 (0.06)	0 [Reference]	.04	
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	

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

[a] 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 (0.41) years. See Table 1 footnote for definition of physical activity levels.

[b] Number of women represents those in the group at baseline.

[c] 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.

Reproduced with permission from: Lee, I. M. et al. JAMA 2010;303:1173-1179.

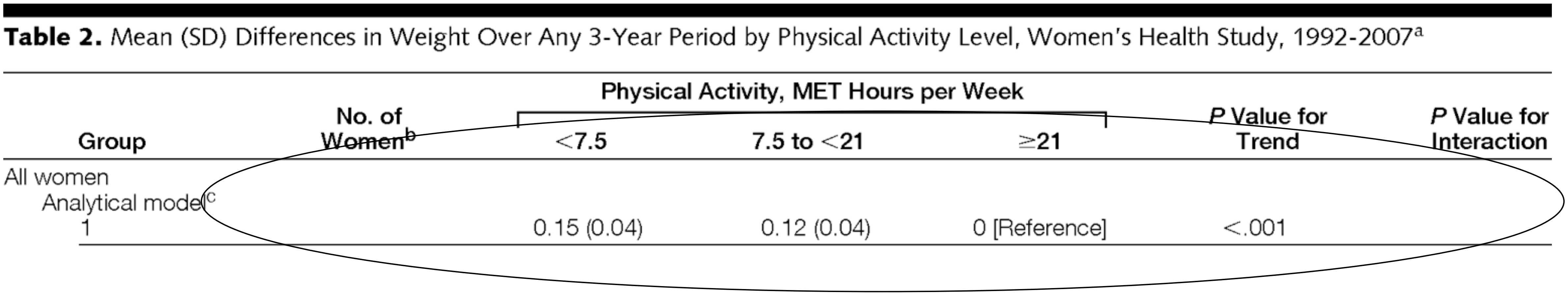


Table 2. Mean (SD) Differences in Weight Over Any 3-Year Period by Physical Activity Level, Women's Health Study, 1992-2007[a]

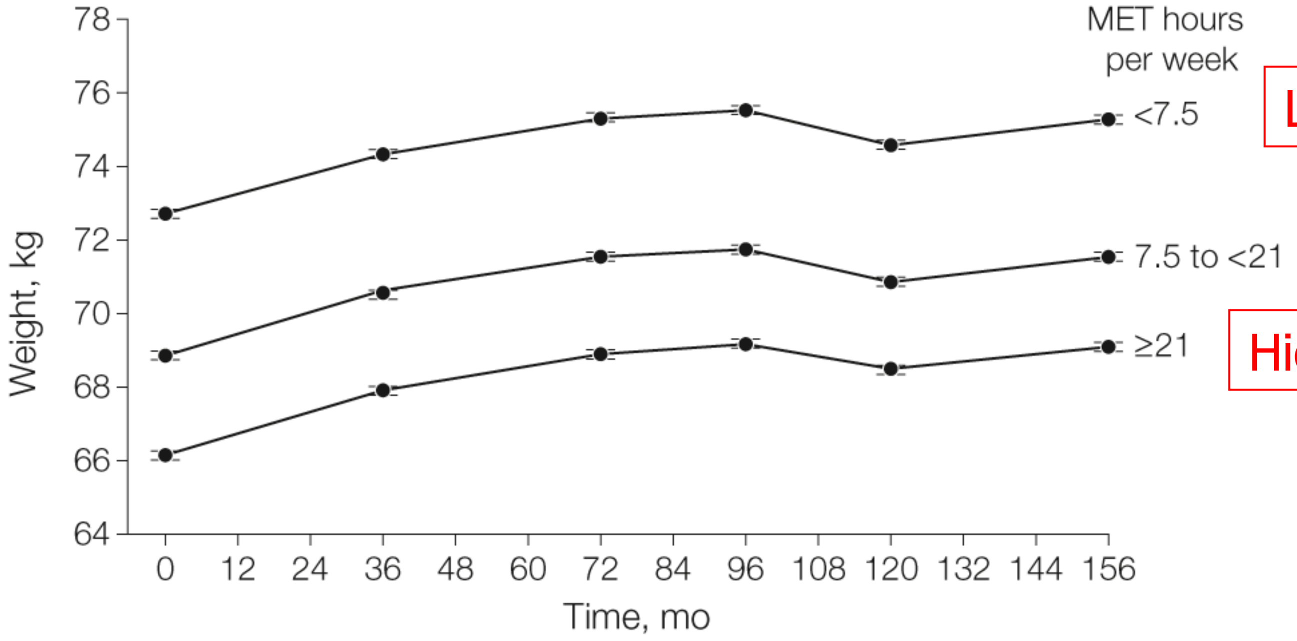
Group	No. of Women[b]	Physical Activity, MET Hours per Week			*P* Value for Trend	*P* Value for Interaction
		<7.5	7.5 to <21	≥21		
All women						
Analytical model[c]						
1		0.15 (0.04)	0.12 (0.04)	0 [Reference]	<.001	

What was the effect size? Those who exercised the least gained 0.15 kg (.33 pounds) more than those who exercised the most over 3 years.

95% confidence interval: 0.09 to 0.44 lbs
per 3 years.

Classic example of a statistically significant effect that is not clinically significant.

A picture is worth...



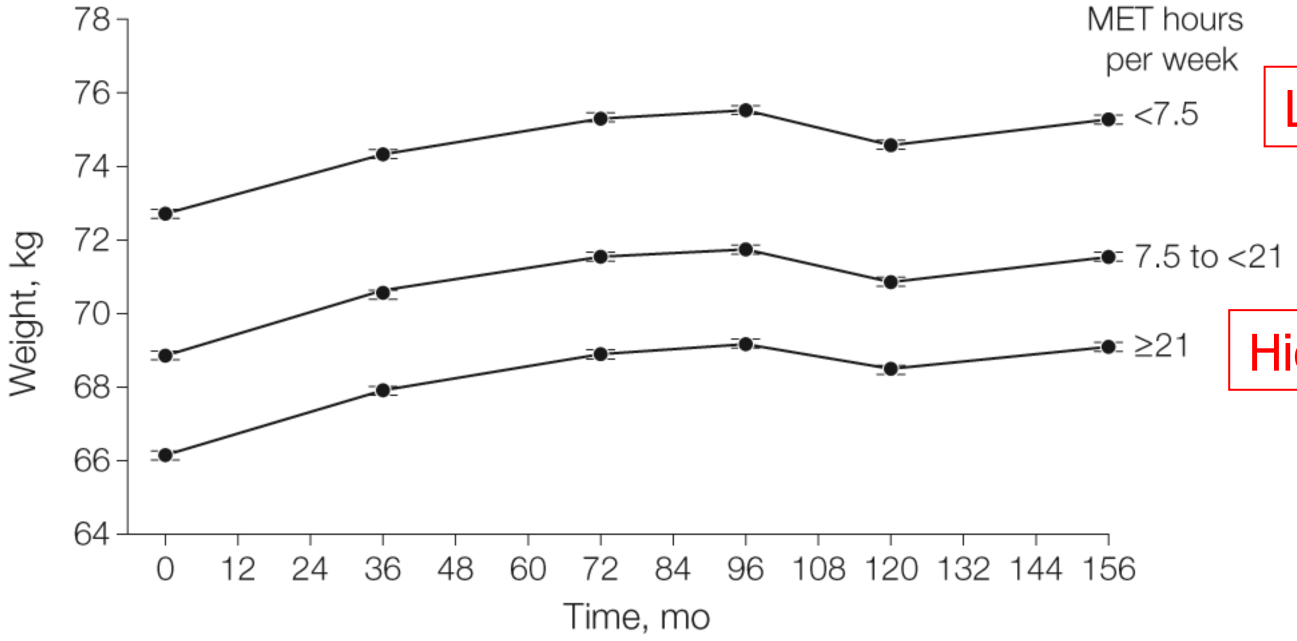
Low exercise group

Medium exercise group

High exercise group

No. by MET hours per week						
<7.5	16 856	15 634	15 153	15 661	13 779	13 353
7.5 to <21	9819	9171	9005	9260	8336	8106
≥21	7404	6924	6808	6992	6264	6107

A picture is worth…



Low exercise group

Medium exercise group

High exercise group

No. by MET hours per week						
<7.5	16 856	15 634	15 153	15 661	13 779	13 353
7.5 to <21	9819	9171	9005	9260	8336	8106
≥21	7404	6924	6808	6992	6264	6107

The heaviest exercisers weigh less to start, *but the weight gain curves between the three baseline groups are almost identical.*

5. Exaggerated effect sizes

Presenting relative risks rather than absolute risks can make effects appear more impressive

Odds ratios (from logistic regression) can distort effects when the outcome is common

Example

Researchers studied the beverage purchases of teenagers at 4 stores; each store was measured at baseline and under 3 "caloric conditions" (signs posted outside the store):

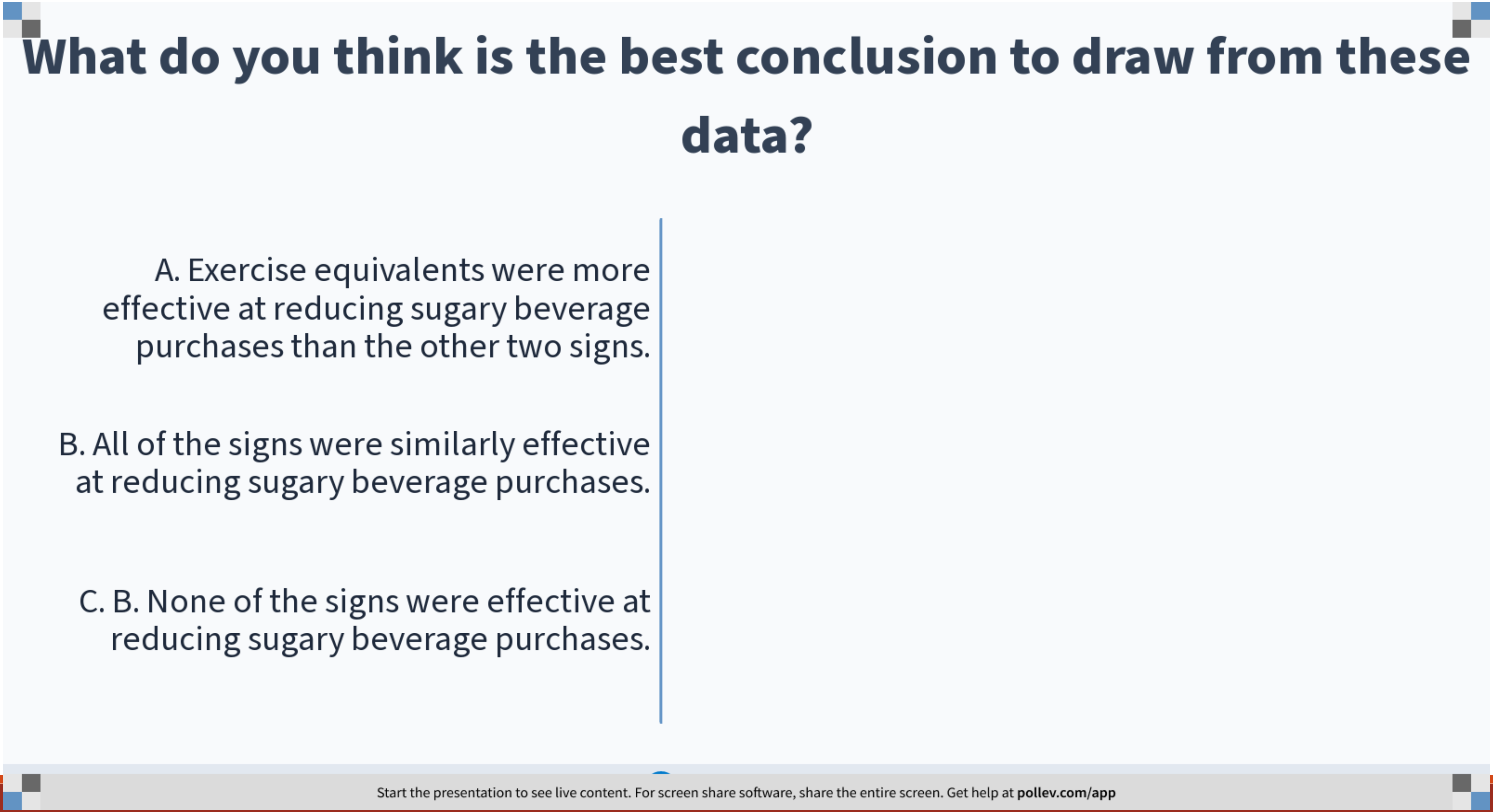
- Absolute calories: "Did you know that a bottle of soda or fruit juice has about 250 calories?"
- Relative calories: "Did you know that a bottle of soda or fruit juice has about 10% of your daily calories?"
- Exercise equivalents: "Did you know that working off a bottle of soda or fruit juice takes about 50 minutes of running?"

Am J Public Health. February 2012, Vol. 102, No. 2, pp. 329-335.

The Results…

Condition	What percent of drinks purchased were sugary beverages?
Pre-intervention (no information)	93.3%
Absolute calories	87.5%
Relative calories	86.5%
Exercise equivalent	86.0%
Any caloric information (overall)	86.7%

What conclusions would you draw?



What do you think is the best conclusion to draw from these data?

A. Exercise equivalents were more effective at reducing sugary beverage purchases than the other two signs.

B. All of the signs were similarly effective at reducing sugary beverage purchases.

C. B. None of the signs were effective at reducing sugary beverage purchases.

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

What do these data tell us?

Posting a sign did reduce sugary beverage purchase: About 6 out of 100 fewer teenagers purchased a sugary beverage. All three messages were similarly effective.

Headlines…

"'Exercise labels' beat out calorie counts in steering consumers away from junk food"

"Researchers: Exercise labels better at keeping teens away from junk food"

"Exercise labels more effective than calorie counts on soda cans"

"If '250 calorie' label doesn't stop you, '50 minute jog' label might"

The authors explaining their results in a university news video:

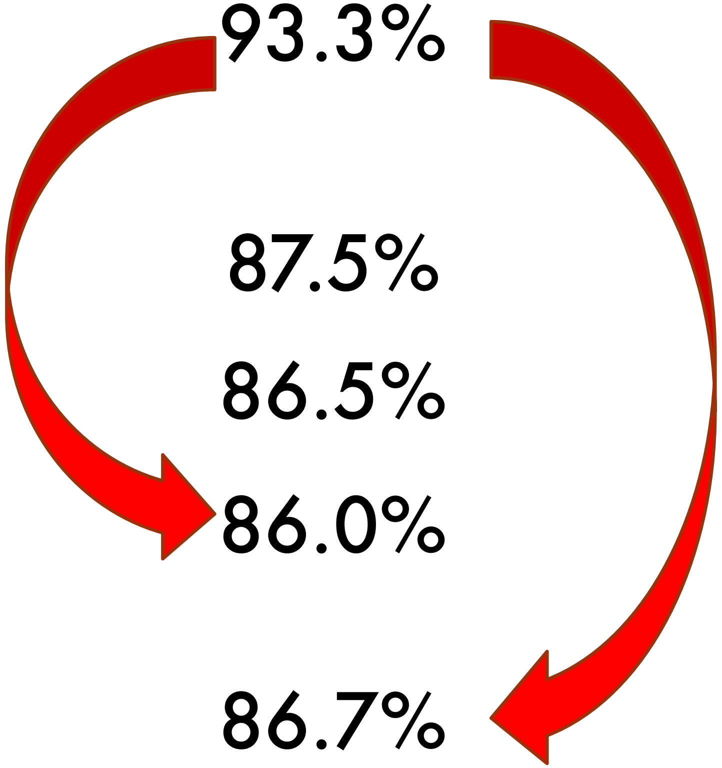
“The results are really encouraging. We found that providing any information (via the three signs) relative to none, ***reduced the likelihood that they would buy a sugary beverage by 40 per cent.***

“Of those three signs, ***the one that was most effective was the physical activity equivalent****.*

“We found that when that sign was posted, ***the likelihood that they would buy a sugary beverage reduced by around 50 per cent.”***

Huh?...

Condition	What percent of drinks purchased were sugary beverages?
Pre-intervention (no information)	93.3%
Absolute calories	87.5%
Relative calories	86.5%
Exercise equivalent	86.0%
Any caloric information (overall)	86.7%



How is this a 50% reduction in risk?

How is this a 40% reduction in risk?

What went wrong? The authors misinterpreted odds ratios.

Condition	Unadjusted Percentage of sugary drinks	Adjusted **Odds ratio**
Pre-intervention (no information)	93.3	1.00 (ref)
Absolute calories	87.5	0.62
Relative calories	86.5	0.59
Exercise equivalent	86.0	0.51
Any caloric information	86.7	0.56

This does not mean a “50% drop in likelihood.

This is not a “40% drop in likelihood.”

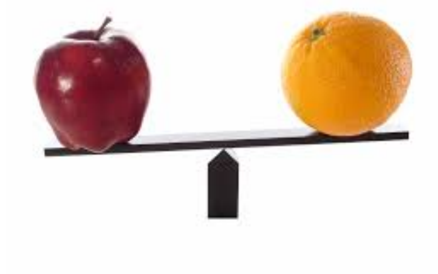
Odds ratio vs. risk ratio…

Risk ratio: $\frac{86\%}{93\%} = 0.92$

Corresponding Odds ratio: $\frac{\frac{86\%}{14\%}}{\frac{93\%}{7\%}} = 0.46$

We cannot interpret the odds ratio as indicating a "54% drop in likelihood"! There is a 54% drop in odds, but only an 8% drop in likelihood (risk)!

6. Wrong comparisons



Authors often steal focus from the main comparison by throwing in p-values from meaningless comparisons.

What do all these statements have in common?

“The effect was significant in the treatment group, but not significant in the control group.”

“Intervention 1 caused a significant change but intervention 2 did not.”

“The effect was significant in subgroup A but not in subgroup B.”

These are all the wrong comparisons!

"The effect was significant in the treatment group, but not significant in the control group."

- Right comparison: treatment vs. control

"Intervention 1 caused a significant change from baseline but intervention 2 did not."

- Right comparison: intervention 1 vs. intervention 2

"The effect was significant in subgroup A but not in subgroup B."

- Right comparison: subgroup A vs. subgroup B

The focus should be on between-group not within-group comparisons.

Exercise labels study again…

Only exercise equivalent labels were significantly different than pre-intervention; odds ratios and 95% confidence intervals:

Absolute calories: OR=0.62 (0.37, 1.04)

Relative calories: OR=0.59 (0.34, 1.02)

Exercise equivalent: OR=0.51 (0.31, 0.85)*

Erroneous Implication: exercise equivalents “beat” absolute and relative calories.

No! The three interventions were statistically indistinguishable.

- Relevant comparison: absolute calories vs. relative calories vs. exercise equivalents

7. Failure to account for corr observations

Correlated observations when pairs or clusters of observations are related and thus are more similar to each other than to other observations in the dataset. Correlated observations require special statistical tests that account for the correlation.

Examples:

The same person measured over time (repeated measures).

Two knees from the same person.

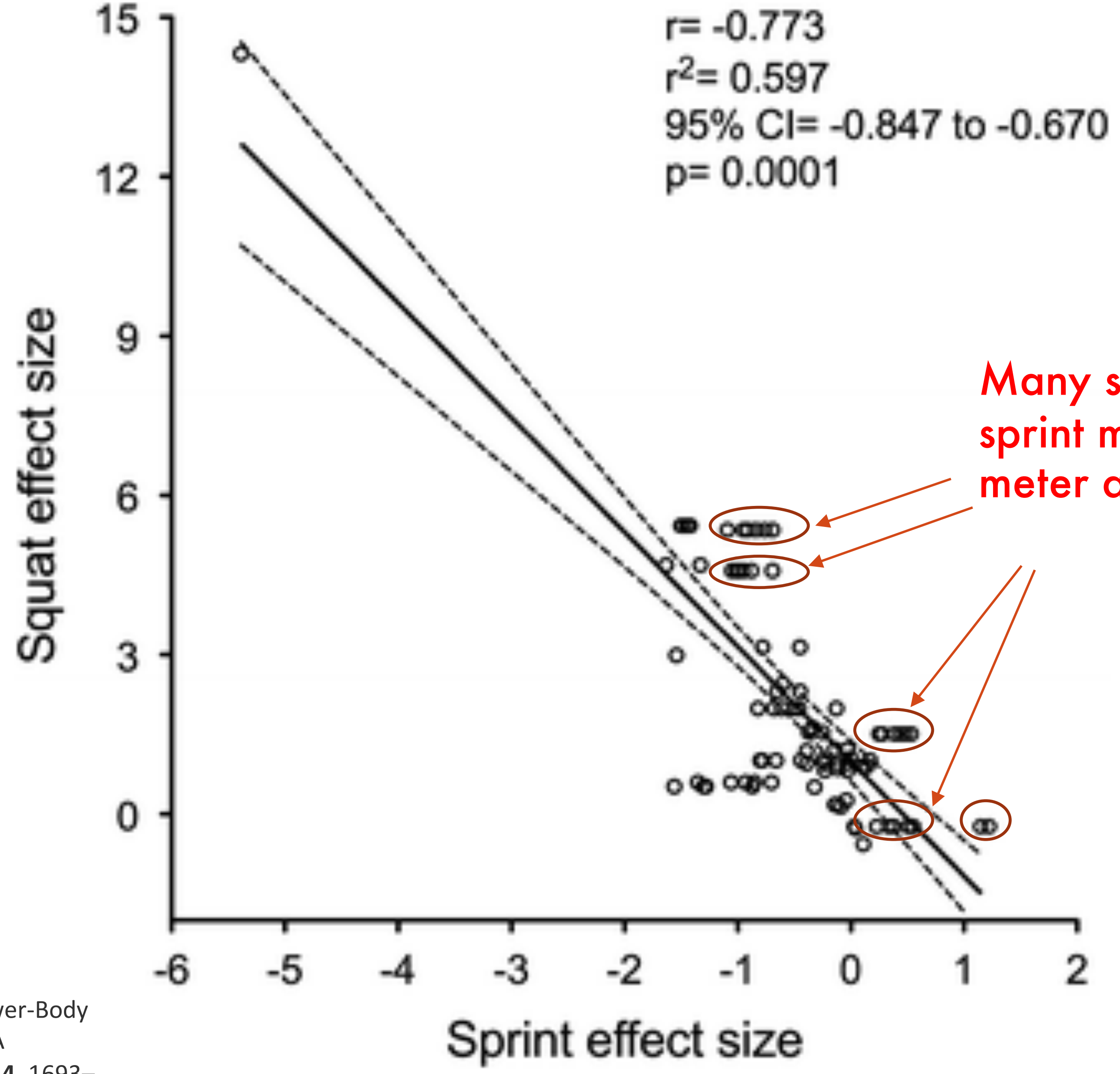
Two hands from the same person.

Related individuals (e.g., twins).

Individuals from the same cluster in a cluster-randomized trial.

Leg strength and sprinting meta-analysis again…

The meta-analysis included 15 studies, but there are 85 data-points included in the correlation analysis.



Many studies included multiple sprint measures (e.g., 10 meter and 30 meter sprints).

The observations from these studies are correlated!

Seitz, L.B., Reyes, A., Tran, T.T. *et al.* Increases in Lower-Body Strength Transfer Positively to Sprint Performance: A Systematic Review with Meta-Analysis. *Sports Med* **44,** 1693–1702 (2014).

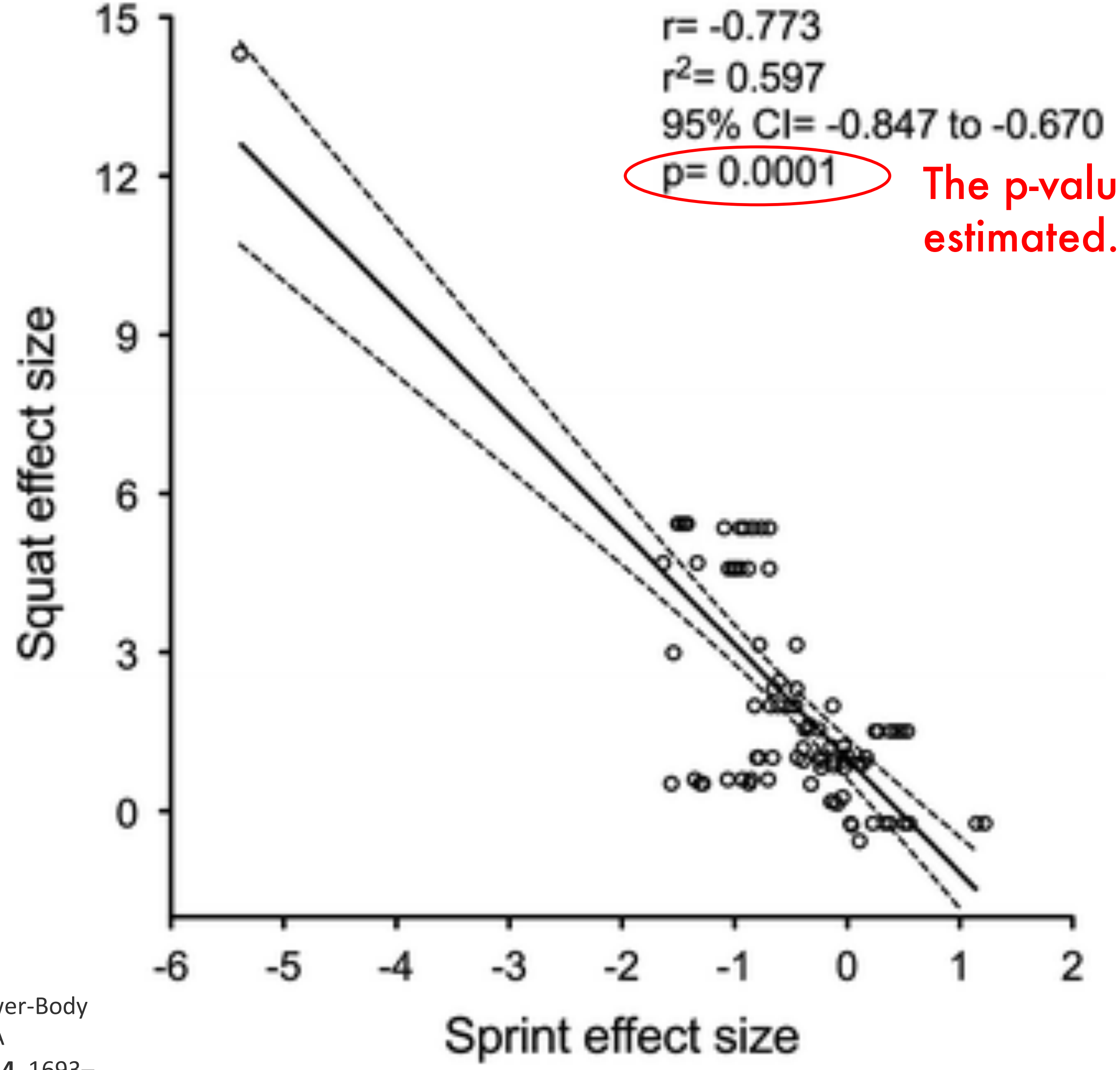
By, ignoring the within-study correlation, the authors have...?

A. Underestimated the p-value.

B. Overestimated the p-value.

C. Correctly estimated the p-value.

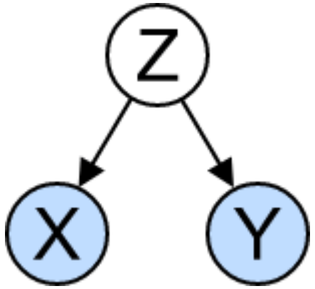
Leg strength and sprinting meta-analysis again…



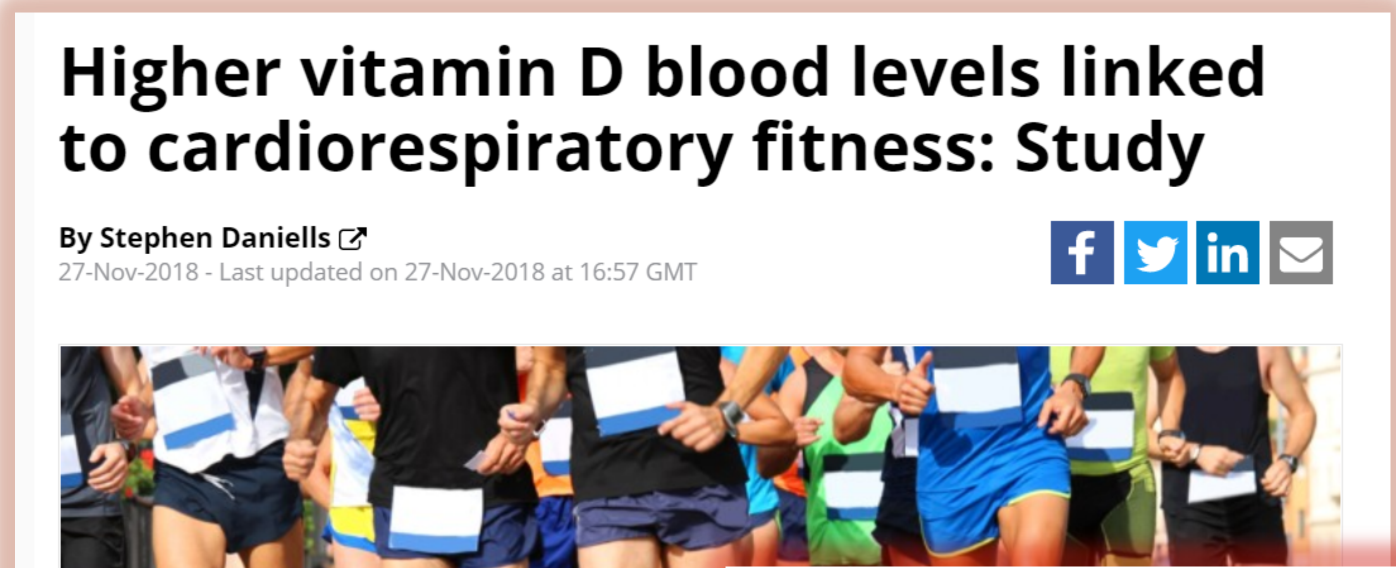
The p-value is under-estimated.

Seitz, L.B., Reyes, A., Tran, T.T. *et al.* Increases in Lower-Body Strength Transfer Positively to Sprint Performance: A Systematic Review with Meta-Analysis. *Sports Med* **44,** 1693–1702 (2014).

8. Residual or unmeasured confounding



“We adjusted for confounders” often doesn’t cut it.

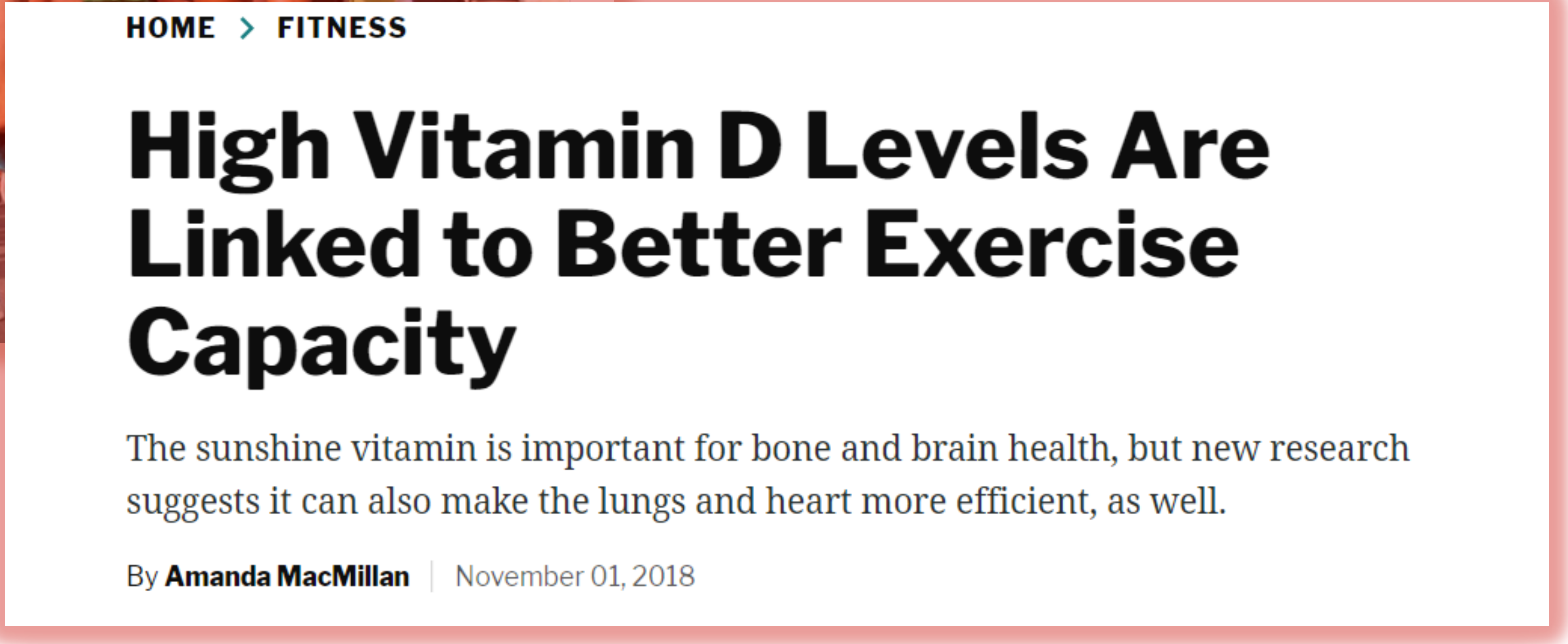


Higher vitamin D blood levels linked to cardiorespiratory fitness: Study

By Stephen Daniells

27-Nov-2018 - Last updated on 27-Nov-2018 at 16:57 GMT

Altmetric Score



HOME › FITNESS

High Vitamin D Levels Are Linked to Better Exercise Capacity

The sunshine vitamin is important for bone and brain health, but new research suggests it can also make the lungs and heart more efficient, as well.

By **Amanda MacMillan** | November 01, 2018

Media Coverage

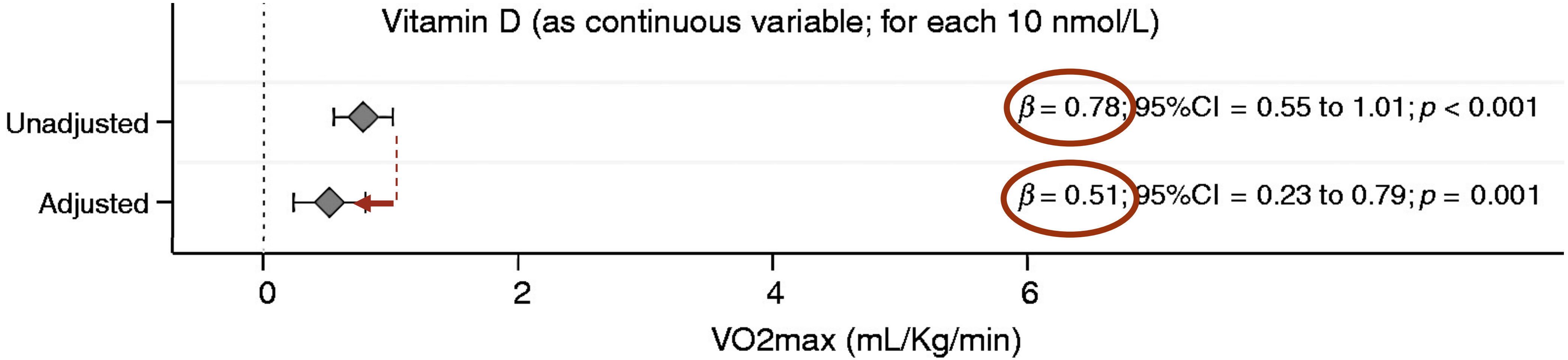
“The study notes that vitamin D could potentially affect cardiorespiratory fitness in several ways. For starters, the nutrient has been shown to boost the production of muscle protein and aid in calcium and phosphorus transport on a cellular level. It may also affect the body’s makeup of fast-twitch muscle fibers, ‘suggesting that vitamin D may improve aerobic fitness,’ the authors wrote.”

https://www.health.com/fitness/vitamin-d-improves-fitness-levels

The Study

- Representative sample of the US population: data from the National Health and Nutrition Survey (NHANES).
- Cross-sectional survey data
- About 2000 participants between the ages of 20 and 49 years
- Examined association between vitamin D levels and VO_2 max

Marawan A, Kurbanova N, Qayyum R. Association between serum vitamin D levels and cardiorespiratory fitness in the adult population of the USA. *European Journal of Preventive Cardiology*. 2019;26(7):750-755. doi:10.1177/2047487318807279

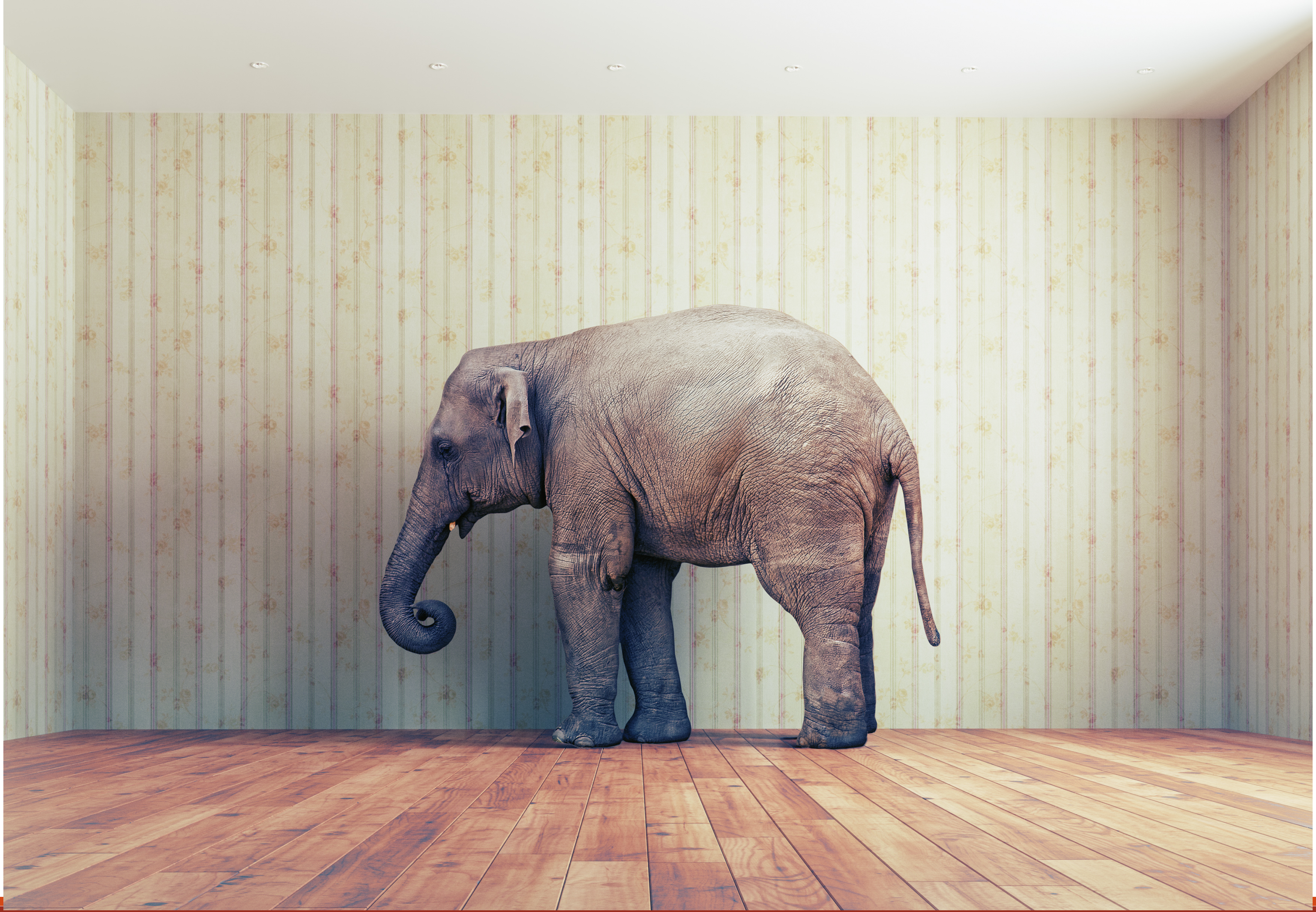


"Linear regression models were adjusted for age, sex, race, BMI, hypertension, diabetes mellitus, smoking, C-reactive protein, total cholesterol, hemoglobin, and renal function as estimated by the GFR."

Marawan A, Kurbanova N, Qayyum R. Association between serum vitamin D levels and cardiorespiratory fitness in the adult population of the USA. *European Journal of Preventive Cardiology*. 2019;26(7):750-755. doi:10.1177/2047487318807279

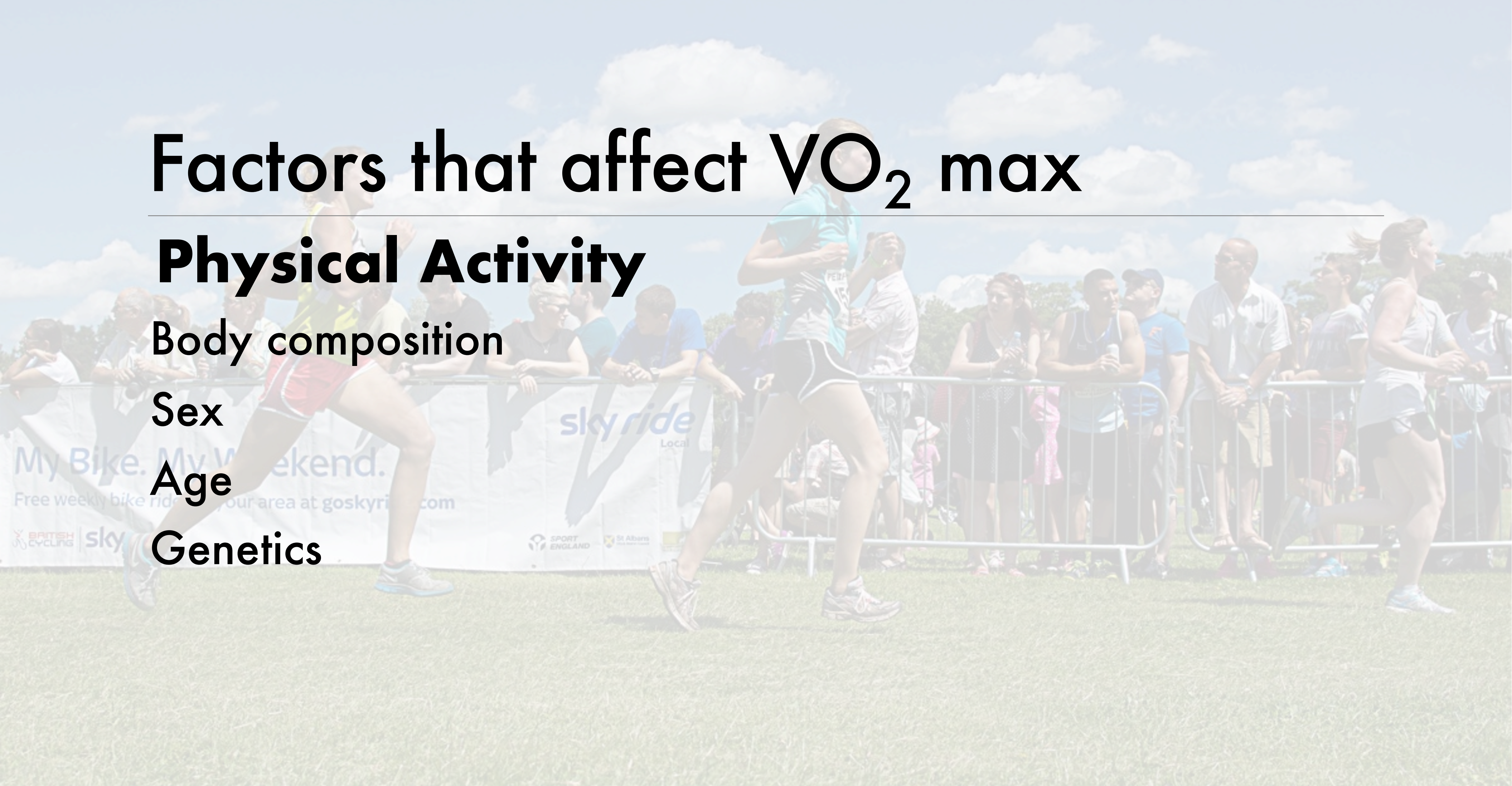
Reaching for biological explanations…

“**Several lines of evidence support the biological plausibility of a potential role of vitamin D in CRF [cardiorespiratory fitness**]. About 3% of all genes are directly or indirectly affected by vitamin D levels and vitamin D receptors are expressed in a large variety of cells, including myocytes.[5,15] **Vitamin D may affect myocytes by increasing muscle protein synthesis and calcium and phosphorus transport in energy production**.[16,17] In addition, **vitamin D may increase the relative number of one type of fast-twitch muscle fibers** (IIa) and decrease another type of fast-twitch muscle fibers (IIb), suggesting that vitamin D may improve aerobic fitness.[18] In addition to its effect on muscles, **animal studies suggest that vitamin D may have a role in heart structure and function**.[5,15,19] Mitochondria from chick cardiomyocytes produced less energy when vitamin D levels were low.[16] In another study, vitamin D deficiency in rodents was associated with decreased myocardial contractility and cardiac output and increased heart rate, changes commonly seen in failing hearts.[20] Vitamin D deficiency has been reported to be associated with decreased myofibrillar area and the increased deposition of myocardial collagen in the extracellular space.[21]”



Factors that affect vitamin D levels

- **Exposure to sunlight**
- Diet
- Supplements
- Obesity
- Skin Pigment
- Age
- Certain diseases
- Genetics
- Season of measurement



Factors that affect VO_2 max

Physical Activity

Body composition

Sex

Age

Genetics

They buried the lead…

Buried in the discussion: "The results were not adjusted for vitamin D intake or physical activity, both of which may have an effect on the observed association."

$p>.05$

9. Misinterpretations of null effects

Lack of statistical significance is not proof of no effect.

NSAIDs study again…

Aim: The purpose of this study was to identify the clinical effectiveness of **oral versus topical NSAIDs** in the treatment of greater trochanteric pain syndrome.

Methods: A retrospective chart review of **25 patients** diagnosed with greater trochanteric pain syndrome were categorized into two groups: those who received **oral etodolac 400 milligrams** twice daily versus **topical diclofenac 3%** one gram two to three times daily for two weeks.

Outcome: Pain scores using the
eline, two-week, and six-week follow-up visits.

Lack of a statistically significant difference is not proof of no effect!

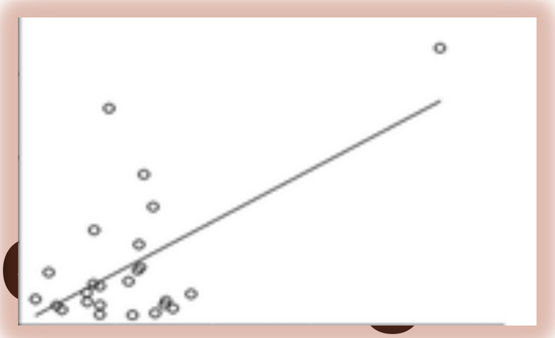
Results: At **two weeks**, there was a statistically significant improvement in pain in both the oral and topical NSAID groups, with no statistically significant difference between the groups (p=0.77).

Similarly, at **six weeks**, there was a statistically significant improvement in pain in both the oral and topical NSAID groups, with no statistically significant difference between the groups (p=0.59).

Conclusion: Based on this study, the use of topical NSAIDs is non-inferior to oral NSAIDs in the treatment of GTPS.

J Pain, 16 (2015), p. s67

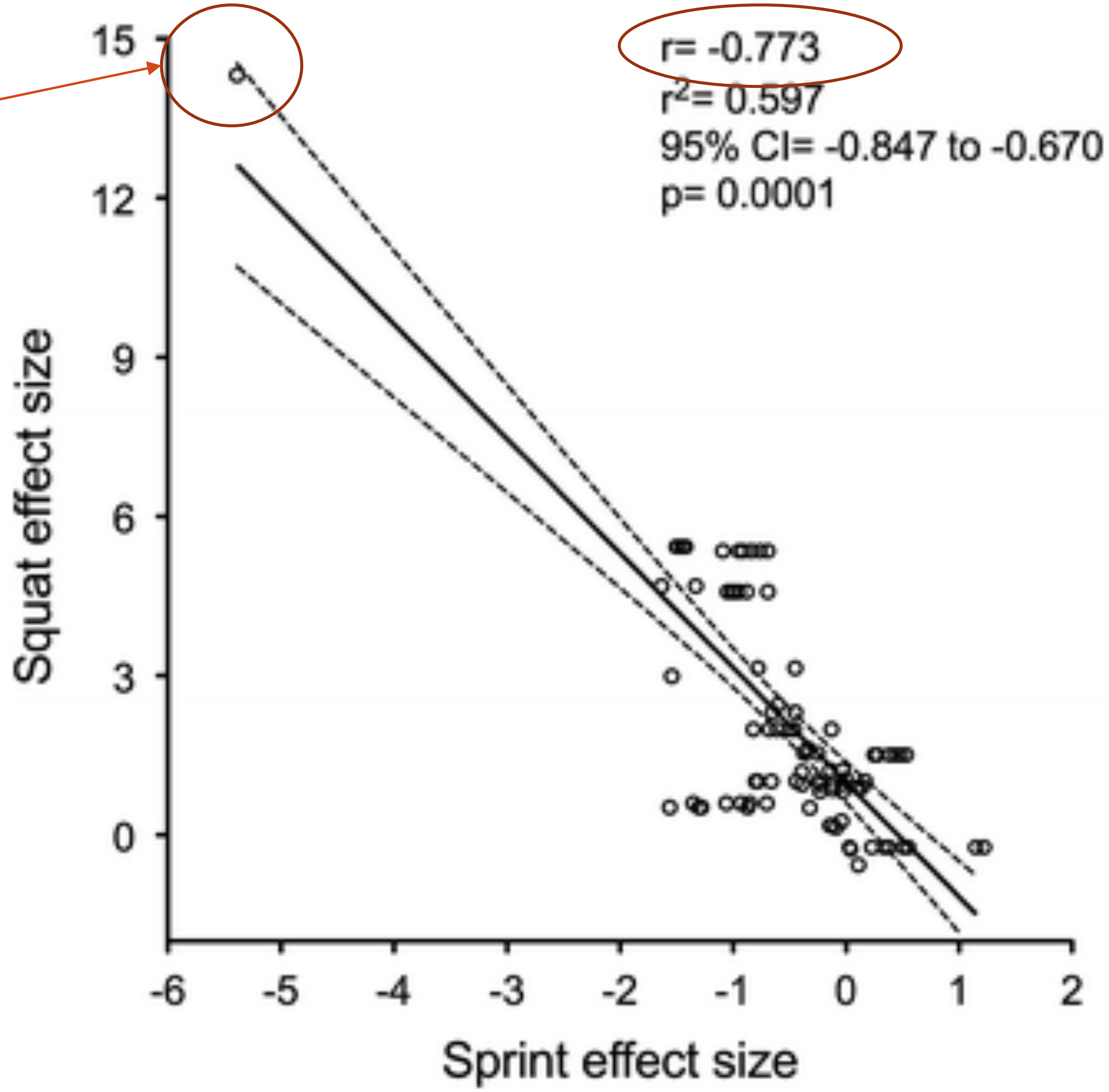
10. Spurious correlations/ov



When sample size is small, a single data point (or a few datapoints) can have undue influence on a correlation coefficient or model.

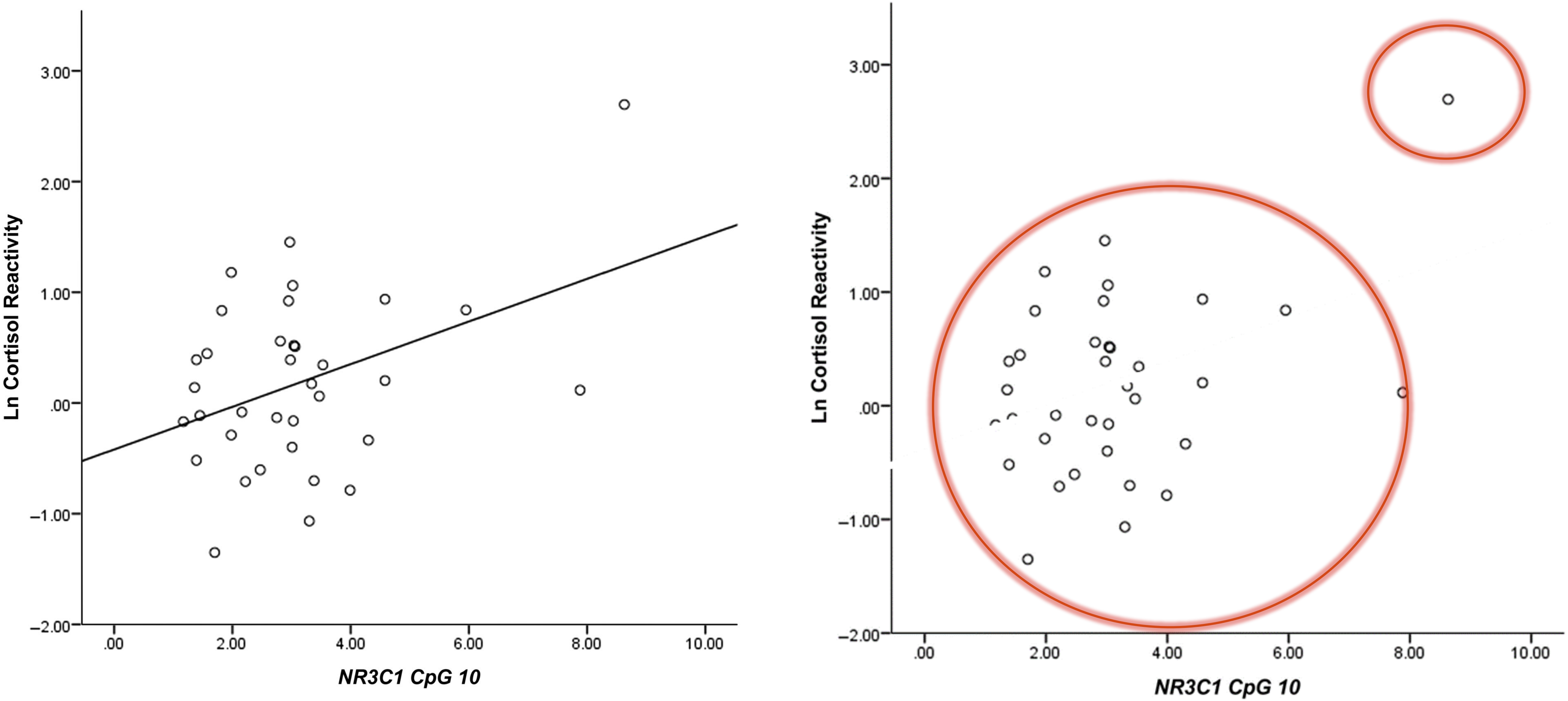
Leg strength and sprinting meta-analysis again…

"Rogue data point"



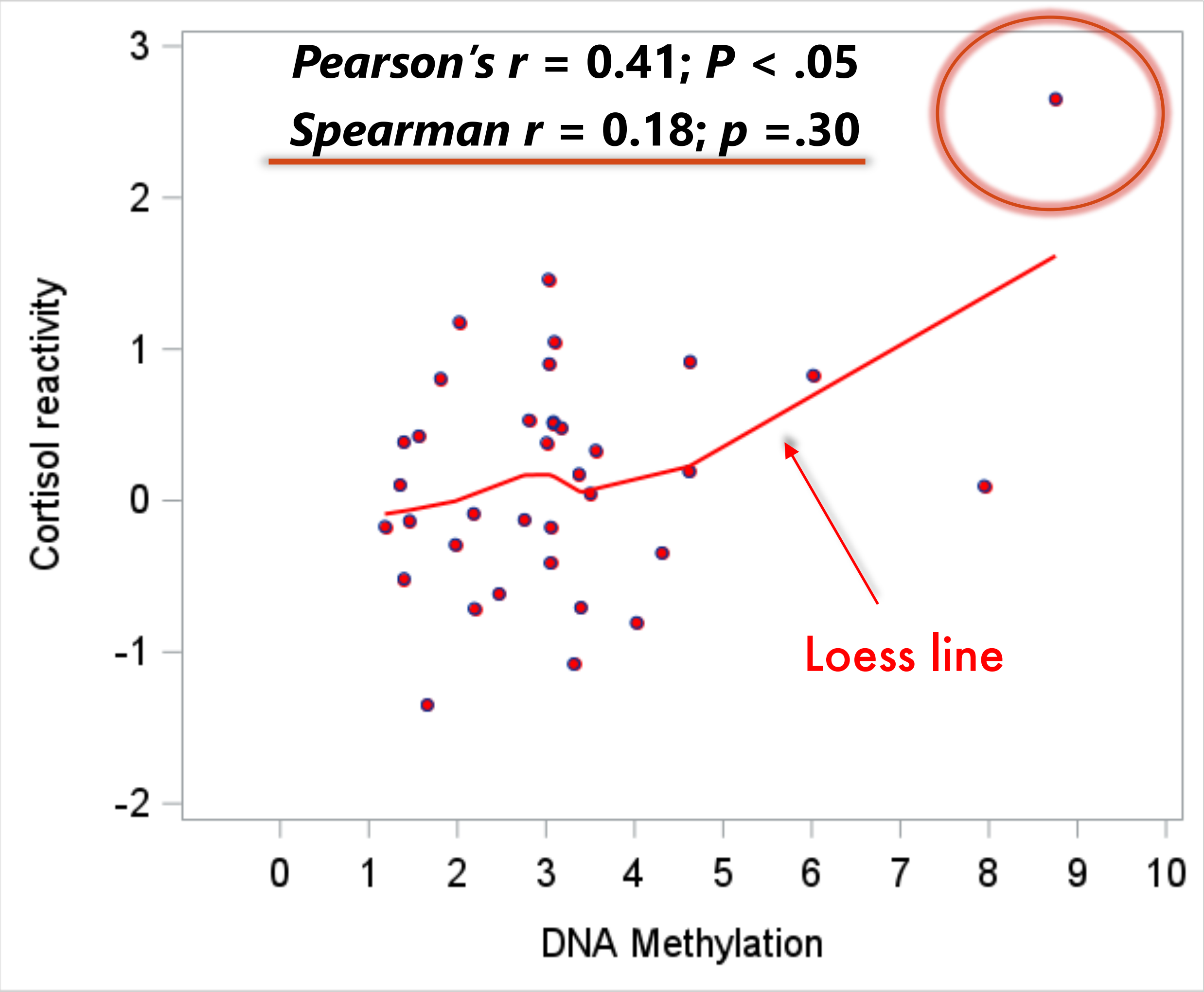
Seitz, L.B., Reyes, A., Tran, T.T. *et al.* Increases in Lower-Body Strength Transfer Positively to Sprint Performance: A Systematic Review with Meta-Analysis. *Sports Med* **44,** 1693–1702 (2014).

Plot showing the correlation between DNA methylation of the glucocorticoid receptor gene (*NR3C1*) and cortisol reactivity, ***r* = 0.41; *P* < .05**



Barry M. Lester, Elisabeth Conradt, Linda L. LaGasse, Edward Z. Tronick, James F. Padbury, Carmen J. Marsit. Epigenetic Programming by Maternal Behavior in the Human Infant. *Pediatrics*, 2018; e20171890 DOI: 10.1542/peds.2017-1890

With influential point

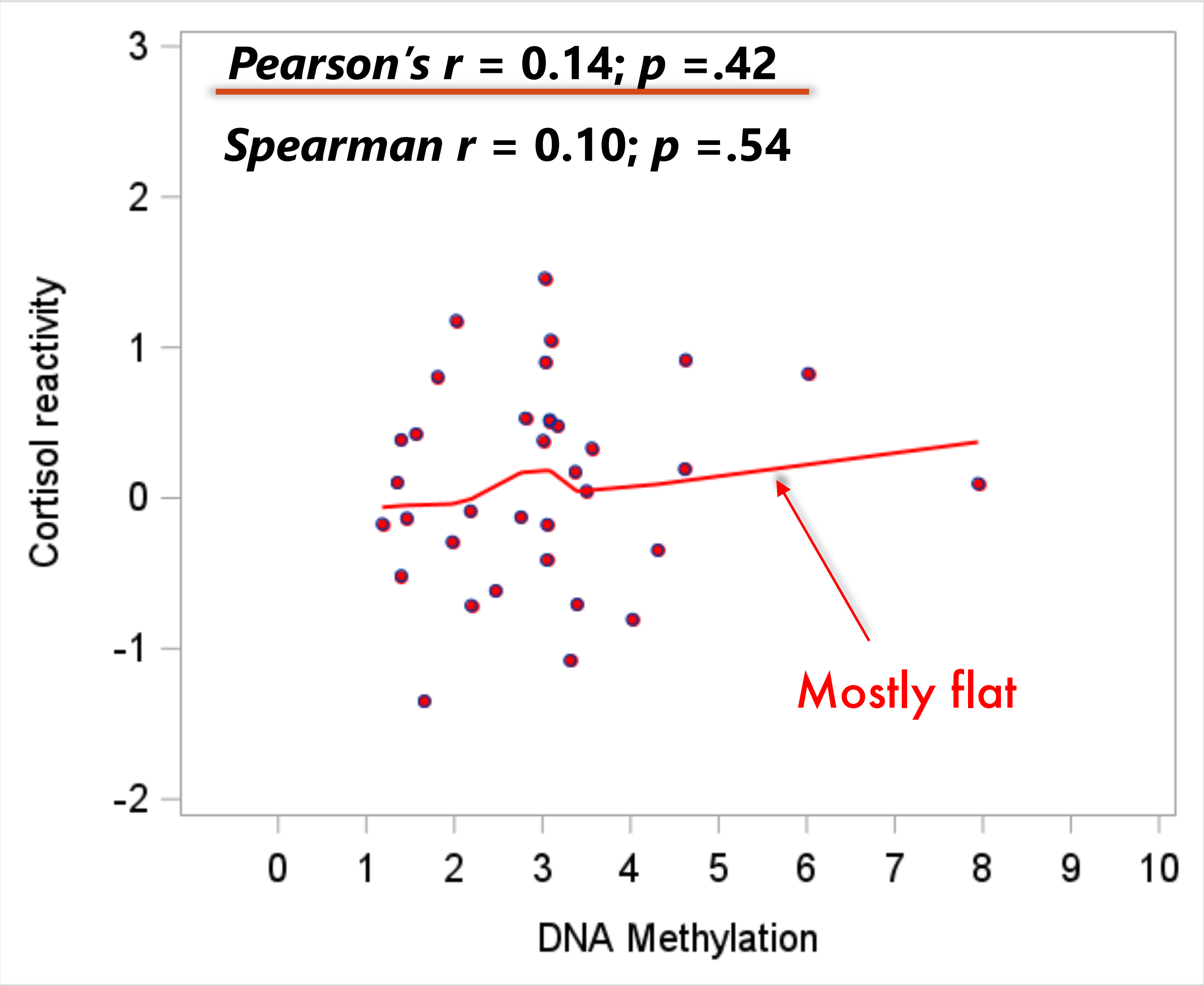


***Pearson's r* = 0.41; *P* < .05**

***Spearman r* = 0.18; *p* =.30**

Loess line

Without influential point



***Pearson's r* = 0.14; *p* =.42**

***Spearman r* = 0.10; *p* =.54**

Mostly flat

Further resources

How to be a Statistical Detective webinar:

https://www.youtube.com/watch?v=JG_gCIGFaQI

Medical Statistics Certificate Program:

https://online.stanford.edu/programs/stanford-medical-statistics-certificate

Statistically Speaking column:

https://onlinelibrary.wiley.com/doi/toc/10.1016/(ISSN)1934-1563.statistics

Relevant PM&R columns:

1. Simple math and data errors

    Avoid Careless Errors: Know your data

    How to be a Statistical Detective

    Ten Common Statistical Errors from All Phases of Research, and Their Fixes

2. Study design mismatched to statistical question

    Ten Common Statistical Errors from All Phases of Research, and Their Fixes

3. Chance findings

    The Problem of Multiple Testing

4. Clinically irrelevant effect sizes

    Clinical Versus Statistical Significance

5. Exaggerated effect sizes

    Understanding Odds Ratios

    Communicating Risks Clearly: Absolute Risk and Number Needed to Treat

Relevant PM&R columns:

6. Wrong comparisons

Misleading Comparisons: The Fallacy of Comparing Statistical Significance

Ten Common Statistical Errors from All Phases of Research, and Their Fixes

7. Failure to account for correlated observations

The Importance of Accounting for Correlated Observations

Ten Common Statistical Errors from All Phases of Research, and Their Fixes

8. Misinterpretations of null effects

Interpreting "Null" Results

9. Residual/unmeasured confounding

The Limitations of Statistical Adjustment

10. Spurious correlations/overfitting

The Value of Scatter Plots

How to be a Statistical Detective

Further resources

How to be a Statistical Detective webinar:

https://www.youtube.com/watch?v=JG_gCIGFaQI

Medical Statistics Certificate Program:

https://online.stanford.edu/programs/stanford-medical-statistics-certificate

Statistically Speaking column:
https://onlinelibrary.wiley.com/doi/toc/10.1016/(ISSN)1934-1563.statistics