Stents: An exploration of design, measurement, analysis, and reporting in clinical research*

Andrew Gelman† and Brahmajee Nallamothu‡

26 Dec 2017

1. Introduction

On 3 Nov 2017, we received an email pointing to a news article entitled, "'Unbelievable': Heart Stents Fail to Ease Chest Pain," along with this note:

> The research design is pretty cool—placebo participants got a fake surgery with no stent implanted. The results show that people with the stent did have better metrics than those with just the placebo . . . but the difference was not statistically significant at 95% confidence, so the authors claim there is no effect! (the difference was significant at 80% confidence). So, underpowered study becomes ammunition in the "stents have no material impact" fight.

It is a well known statistical fallacy to take a result that is not statistically significant and report it as zero. Here a study was performed and yielded $p = 0.20$ and a null result. Had the data happened to come out with $p = 0.04$, would the headline have been, "'Believable': Heart Stents Indeed Ease Chest Pain"? A lot of certainty seems to be hanging on a small bit of data.

An examination of lack of statistical significance of this finding leads to larger questions about science reporting, medical research, and statistically-based decision making.

2. The reporting

Al-Lamee et al. (2017) recently published a randomized control trial of percutaneous coronary intervention (stents) for relief of symptoms for 200 patients with angina (chest pain from blocked arteries). The study, called ORBITA (Objective Randomised Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina), was notable for being a blinded experiment in which half the patients received stents and half received a placebo procedure in which a sham operation was conducted, so that the patients and subsequent caregivers did not know which treatment they received. In follow-up, patients were asked to guess their treatment: 24% said they did not know, and of the 76% who were willing to guess, 56% guessed correctly and 44% guessed wrong, which is within the range that could be expected by chance alone.

The summary result from the study, as summarized by Al-Lamee in a press release (O'Hare, 2017): "even though the stents improved blood supply, they didn't provide more relief of symptoms compared to drug treatments, at least in this patient group."

The finding was reported by Kolata (2017) in the *New York Times* as "unbelievable . . . stunned leading cardiologists by countering decades of clinical experience. . . . The idea that stenting relieves chest pain is so ingrained that some experts said they expect most doctors will continue with stenting." Indeed, one of us was quoted there as being humbled by the finding, as cardiologists who

*We thank Doug Helmreich for bringing this example to our attention, John Carlin, Avi Feller, Brian Goodman, Frank Harrell, and Shira Mitchell for helpful comments, and the Office of Naval Research and Defense Advanced Research Project Agency for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York

‡Division of Cardiovascular Diseases and Department of Internal Medicine, University of Michigan, Ann Arbor

put stents in had expected a positive result. On the other hand, Kolata also noted that "there have long been questions about their effectiveness." And the very willingness of doctors and patients to participate in a controlled trial with a sham treatment suggests that there must have been some existing skepticism about the use of stents for angina patients.

Indeed, Belluz (2017) led off her news article in *Vox* with: "There's an epidemic of unnecessary medical treatments . . . One of the prime examples of a dubious treatment . . . is cardiologists putting little mesh tubes called stents in patients with stable angina . . . The idea is that stents should help soothe the suffering of patients with angina (or chest pain) and drive down the risk of a heart attack and death in the future. But studies show that stable angina can be well controlled with medication. And researchers have found that stenting chest pain patients doesn't help them live longer or reduce their risk of disease—in fact, heart attacks and strokes can be potentially deadly side effects of stent procedures. There's also been a lingering question about whether stents truly work to relieve pain."

So, just at the most basic level, the recent study is an unbelievable game-changer (according to some sources) or confirmation of expert opinion (according to others). The research article and the news reports present the study as revealing that stents' pain-relieving effects could be due to placebo—but it is not clear how much of a placebo effect we should expect in a study in which patients know that they only have a 50% chance of receiving the treatment.

Here, though, we want to consider the statistical evidence, in particular the claim that stents did not, in the words of Al-Lamee et al. (2017), "increase exercise time by more than the effect of a placebo procedure." As noted in Section 1, this claim of no difference is not literally correct: exercise time, the primary outcome in the preregistered design, increased on average by 28.4 seconds in the treatment group, compared to an increase of only 11.8 seconds in the control group. This difference was not, however, statistically significant, hence following the conventional rules of scientific reporting it was treated as zero, thus an instance of the regrettably common statistical fallacy of presenting a non-statistically-significant results as confirmation of the null hypothesis (Harrell, 2017a). We return to the question of whether changes of this magnitude should be considered clinically important.

3. Improving the data analysis by adjusting for pre-treatment imbalance

So, what about that non-significant result? Figure 1 summarizes the results for eight measured outcomes from the stents experiment, with the primary outcome, shown as the top row of data. Average exercise time increased from 528.0 to 556.3 in the treatment group and from 490.0 to 501.8 in the control group. The average difference in gain scores was 16.6 with a standard error of 9.8 (computed by taking the length of the reported 95% interval and dividing by 4), yielding a p-value of 0.20.

But the estimate using gain scores does not make full use of the data, as discussed, for example, by Vickers and Altman (2001), Gelman and Hill (2007), and Harrell (2017a). Here's the problem: As can be seen in Figure 1, the treatment and placebo groups differ in their pre-treatment levels of exercise time, with mean levels of 528.0 and 490.0, respectively. This sort of difference is fine—randomization assures balance only in expectation—but is important to adjust for this discrepancy in estimating the treatment effect. In the published paper, the adjustment was performed using the estimate,

$$\text{gain score:} \quad (\overline{y_{\text{post}} - y_{\text{pre}}})^T - (\overline{y_{\text{post}} - y_{\text{pre}}})^C, \tag{1}$$

but this *over-corrects* for differences in pre-test scores. It is the familiar problem of "regression to the mean": just from natural variation, we would expect the people with lower scores to improve,

Measurement	Treatment				Control				Comparison	
		Pre	Post	Gain		Pre	Post	Gain		
	N	$\bar{y}$ (sd)	$\bar{y}$ (sd)	diff (ci)	N	$\bar{y}$ (sd)	$\bar{y}$ (sd)	diff (ci)	est (ci)	p
Exercise time (seconds)	104	528.0 (178.7)	556.3 (178.7)	28.4 (11.6, 45.1)	90	490.0 (195.0)	501.8 (190.9)	11.8 (−7.8, 31.3)	16.6 (−8.9, 42.0)	0.200
Peak oxygen uptake (mL/min)	99	1715.0 (638.1)	1713.0 (583.7)	−2.0 (−54.1, 50.1)	89	1707.4 (567.0)	1718.3 (550.4)	10.9 (−47.2, 69.0)	−12.9 (−90.2, 64.3)	0.741
SAQ-physical limitation	100	71.3 (22.5)	78.6 (24.0)	7.4 (3.5, 11.3)	88	69.1 (24.7)	74.1 (24.7)	5.0 (0.5, 9.5)	2.4 (−3.5, 8.3)	0.420
SAQ-angina frequency	103	79.0 (25.5)	93.0 (26.8)	14.0 (9.0, 18.9)	90	75.0 (31.4)	84.6 (27.7)	9.6 (3.6, 15.5)	4.4	0.260
SAQ-angina stability	102	64.7 (25.5)	60.5 (23.7)	−4.2 (−10.7, 2.4)	89	68.5 (24.3)	63.5 (25.6)	−5.1 (−11.7, 1.6)	0.9 (−8.4, 10.2)	0.851
EQ-5D-5L QOL	103	0.80 (0.21)	0.83 (0.21)	0.03 (0.00, 0.06)	89	0.79 (0.22)	0.82 (0.20)	0.03 (0.00, 0.07)	0.00 (−0.04, 0.04)	0.994
Peak stress wall motion index score	80	1.11 (0.18)	1.03 (0.06)	−0.08 (−0.11, −0.04)	57	1.11 (0.18)	1.13 (0.19)	0.02 (0.03, 0.06)	−0.09 (−0.15, −0.04)	0.0011
Duke treadmill score	104	4.24 (4.82)	5.46 (4.79)	1.22 (0.37, 2.07)	90	4.18 (4.65)	4.28 (4.98)	0.10 (−0.99, 1.19)	1.12 (−0.232.47)	0.104

Figure 1: *Summary data comparing stents to placebo, from Table 3 of Al-Lamee et al. (2017).*

relative to the average, and the people with higher scores to regress downward. The optimal linear estimate of the treatment effect is actually,

$$\text{regression estimate:} \quad (\overline{y_{\text{post}} - \beta\, y_{\text{pre}}})^T - (\overline{y_{\text{post}} - \beta\, y_{\text{pre}}})^C, \tag{2}$$

where β is the coefficient of y_{pre} in a regression of y_{post}, also controlling for the treatment indicator. The gain score (1) is a special case of the regression estimate (2) corresponding to $\beta = 1$. Given that the pre-test and post-test measurements have nearly identical variances (as can be seen in Figure 1, we can anticipate that the optimal β will be less than 1, which will reduce the correction for difference in pre-test and thus increase the estimated treatment effect, and also slightly decrease the standard error. As a result, the improved analysis of these data should lead to a lower p-value.

When we realized this, we got excited. After all, the data already show a difference, and the p-value of 0.20 is already not so far from statistical significance. If a cleaner analysis of the data could improve the result enough to bring p below 0.05, what a story it would be! The reanalysis would entirely change the reporting of the experiment, spinning the story from a failure of stents to a success, along with an advertisement for the value of some very basic statistics.

But how to perform the reanalysis, given that the data were not immediately available? It turns out that the information required to perform the regression is right there in Figure 1! For each of treatment and control, we are given the standard deviation of the pre-test measurements, the standard deviation of the post-test measurements, and the standard deviation of their difference, which can be obtained by taking the width of the confidence interval for the difference, dividing by 4 to get the standard error of the difference, and then multiplying by $\sqrt{n}$ to back out the standard deviation. Then we use the rule, $\text{sd}(y_2 - y_1) = \sqrt{\text{sd}(y_1)^2 + \text{sd}(y_2)^2 - 2\rho\,\text{sd}(y_1)\text{sd}(y_2)}$ and solve for ρ, the correlation between before and after measurements within each group. The result in this case is $\rho = 0.88$ within each group. We then convert the correlation to a regression coefficient of y_2 on y_1 using the well-known formula, $\beta = \rho\,\text{sd}(y_2)/\text{sd}(y_1)$, which yields 0.88 for the treated and 0.86 for the control group.

We use the average, $\beta = 0.87$, in (2) and get an estimate of 21.3 (indeed, quite a bit higher

than the reported difference in gain scores of 16.6) with a standard error of 12.5 (very slightly lower than 12.7, the standard error of the difference in gain scores). But the estimate is not quite two standard errors away from zero: the z-score is 1.7, and the p-value is 0.09.

4. By reporting the results differently, the data can be presented as a success or a failure of stents

Damn! So close, yet not quite there. Our anticipated blockbuster story—Statistical Sleuth Reanalyzes Data, Turns a Reported Null Effect into a Statistically Significant Effect—did not quite materialize.

And yet . . . had it been so desired, this experiment could have been presented as positive evidence in favor of stents. In some settings, a p-value of 0.09 would be considered statistically significant; for example, in a recent social science experiment published in the *Proceedings of the National Academy of Sciences*, Sands (2017) presented a causal effect based on $p < 0.10$, and this was enough for publication in a top journal and in the popular press, as this work was mentioned uncritically by Resnick (2017) without any concern regarding significance levels—and that was in *Vox*, the same media outlet that reported the stents experiment as showing a null effect. Had Al-Lamee et al. performed the appropriate statistical analysis with their data and chosen to publish in *PNAS* rather than *Lancet*, they could indeed have confidently reported a causal effect of stents.

It is possible that further precision could be gained by controlling for imbalances in age, smoking status, and other pre-treatment variables, but given the high correlation between pre-test and post-test exercise time measurements, we doubt such additional regression adjustments would do much, and in any case we do not presently have access to the data that would be required to perform the necessary calculations.

When thinking of how the results could be reported, however, we can back up and consider some simpler options. For example, one could just take the summaries from Figure 1 and report them as follows: With stents, there is a statistically significant improvement of 28.4 seconds (with a 95% confidence interval that clearly excludes zero); with placebo, there is no statistically significant change. Thus, the data show that stents work and placebo has no effect. Such a conclusion would be inappropriate, as it would be making the error of comparing significance to non-significance (see Gelman and Stern, 2006). But this error appears in published papers all the time, including in top journals, so it represents another way the data could have been reported.

Another set of choices in the data analysis involves the decision of what to do with the eight measures shown in Figure 1, along with a few other outcomes reported by Al-Lamee et al. (2017) but not reported here. Most of the gain scores in Figure 1 appear to be small (a difference of 16.6 in a scale where the average is about 500; a difference of 12.9 compared to an average of 1700; a difference of 2.4 compared to an average of 70; etc.) and most are not statistically significant; indeed the p-values are not far from uniformly distributed between 0 and 1. On the other hand, if the outcomes are moderately correlated, then some sort of combined score—even a simple average of normalized outcomes for each person—could reveal a clear improvement of treatment compared to control. Other analysis options include the decisions of how to handle the (relatively small amount of) missing data in the experiment.

Our point here is not at all to criticize Al-Lamee el al. for reverse "p-hacking" (Simmons, Nelson, and Simonsohn, 2011)—looking for a null result. Rather we wish to emphasize the flexibility inherent both in data analysis and in reporting, even in the case of a clean randomized study. We are pointing out the potential fragility of the stents-didn't-work story in this case. The existing data could easily have been presented as a success for stents compared to placebo by authors who were aiming for that narrative. And with just a slight change in the data, the results could've been

seen to have been unambiguously in favor of stents. To demonstrate this, we performed a simple bootstrap analysis, computing the results that would have been obtained from reanalyzing the data 1000 times, each time resampling patients from the existing experiment with replacement (Efron, 1979). We could not quite do this, as the raw data were not available to us, so we approximated using the normal distribution based on the observed z-score of 1.7. The result was that, in 40% of the simulations, stents outperformed placebo at a statistically significant level. This is not to say that stents really are better than placebo—the data also appear consistent with a null effect—just that the take-home point of this experiment could easily have gone the other way, for many different reasons.

5. Clinical significance

From Al-Lamee et al. (2017): "Evidence from placebo-controlled randomised controlled trials shows that single antianginal therapies provide improvements in exercise time of 48–55 s. ... Given the previous evidence, ORBITA [the new study] was conservatively designed to be able to detect an effect size of 30 s." Given the general overestimation of treatment effects in the presence of selection for statistical significance (see, for example, Gelman and Carlin, 2014), one would expect published estimates to be too high, and in any case the prior range of 48–55 seems much too narrow. The estimated effect of 21 seconds with a standard error of 12 seconds is, however, consistent with the "conservative" effect size estimate of 30 seconds given in the published article. So, yes, the experimental results are consistent with a null effect, and they are even more consistent with a small positive effect.

One might ask about the clinical significance of any treatment effect. Suppose we take the point estimate from the data at face value. How should we think about an increase in average exercise time of 21 seconds, given that it is relative to a baseline of 500 seconds? One way to conceptualize this is in terms of percentiles. The data show a pre-randomization distribution (averaging the treatment and control groups) with a mean of 509 and a standard deviation of 188. Assuming a normal approximation, an increase in exercise time of 21 seconds from 509 to 530 would take a patient from the 50th percentile to the 54th percentile of the distribution. Looked at that way, it would be hard to get excited about this effect size, even if it were statistically significant.

But there were other signals from the ORBITA study that seemed to suggest improvements in ischemia with PCI as assessed by other endpoints ("PCI did significantly reduce ischemia as assessed by FFR, iFR and stress echo," American College of Cardiology, 2017). Without the longitudinal data to observe the outcomes that matter most to patients—will they live longer and have fewer heart attacks—much remains uncertain.

The larger question has to be the balancing of the long-term benefits of stents on a patient with the risk of the operation. It does not seem reasonable for a person to risk life and health by submitting to a surgical procedure, just for a potential benefit of 21 seconds of exercise time—or even the hypothesized larger benefit of 50 seconds, which would still only represent a 10% improvement for an average patient in this study. Maybe a 10% increase is consequential in this case: perhaps this small gain in exercise time is associated with a much greater feeling of health and well-being. But, if so, we would think that this larger gain would have been apparent in the assessments of angina frequency or severity, and it was not.

Part of the biggest concern here is that these patients were already doing pretty well on medications—that is, they had a low symptom burden for angina after optimizing their drugs as much as possible. One of the great debate points in this whole issue is that those who want to discount ORBITA suggest that the patients that it enrolled reflect only a small number encountered in routine practice, while those who argue that ORBITA is transformative say that this group

makes up most patients undergoing stents for stable coronary artery disease.

Another clue may be provided from this remark from a doctor quoted by Kolata (2017), that stents could be given to patients who "don't want drugs or can't take them." So the relevant decision may be stents plus some drugs vs. more drugs, not stents vs. nothing. And drugs can have risks too.

What about longer-term benefits, which we would think would be the overwhelming concern for heart patients? According to the above-cited press release and news articles, stents are acknowledged to be effective after heart attacks but not for patients with stable angina. So, from that clinical perspective, the question is not, "Are stents better than placebo?", but ultimately "Who should get stents?" or "When should stents be considered for a patient?"

Are stents really being given to stable angina patients just to reduce pain and improve short-term fitness? Or is there a belief that stents have long-term benefits for these patients as well, despite the earlier COURAGE study that, according to Al-Lamee et al., found "no difference in myocardial infarction and death rates between patients with stable coronary artery disease who underwent PCI and controls"? This would seem to be the key question, in which case the short-term effects, or lack thereof, found in the ORBITA study are close to irrelevant. Other larger trials such as ISCHEMIA are considering this more fundamental question but will not have a placebo procedure.

6. Discussion

The search for better medical care is an incremental process, with lots of incomplete evidence accumulating, and there is a fundamental incompatibility between that idea and the up-or-down reporting of individual studies based on statistical significance. At this point it's not clear how to think about the relevance of this recent experiment to interventional cardiology practice despite its novel and provocative study design.

A reanalysis of the summary data from Al-Lamee et al. (2017) reveals a stronger estimated effect that is closer to the conventional boundary of statistical significance, indicating that the study could easily have been reported as evidence in favor of, rather than against, the effectiveness of stents for angina patients. And, from our brief flurry of excitement over the possibility that a simple reanalysis could change the significance level, we are again reminded of the sensitivity of headline conclusions to decisions in data processing. In any case, though, the observed increases in exercise time, even if statistically significant, do not appear at first glance to be of much clinical importance, compared to the much more relevant long-term health outcomes.

In the design, evaluation, and reporting of experimental studies, there is a norm of focusing on the statistical significance of some particular outcome—in this case, change in average exercise time. The resulting conclusions will be fragile: unless the underlying effect is huge, p-values are extremely noisy, as demonstrated at the end of Section 4.. An experiment may be designed to have 80% power but this is typically conditional on an overestimated effect size (see Gelman, 2018), and in any case does not address the important question of variation in treatment effects.

In the stents example, there seems to be a disconnect between the findings emphasized in the recent study—however presented—and the larger context of treatments for heart disease. From a statistical perspective, this may indicate a problem with the framing of clinical trials, including this one, as attempts to discover whether a treatment has a statistically significant effect. Power calculations are used in an attempt to assure stable estimates and a good chance of the experiment being successful, but within these constraints there can be a push toward convenience rather than relevance of outcome measures. The ORBITA study shows us the potential problems this can raise when a finding is close but not close enough.

The ORBITA study was never meant to be definitive—it was designed to find a physiological effect of stenting on exercise time. If the result had been unambiguously positive, people would have still asked for a full outcomes trial to be done, and some would have suggested that the study was pointless as the difference was not clinically significant. A possible reason why the study was limited in its size and design of these surrogate outcomes was because this is all that could have passed an ethical board given the novelty of the placebo procedure. Further background on the study from Darrel Francis, the senior author on the project, appears at Harrell (2017b). Beyond any immediate news reports, one impact of ORBITA is that a bigger placebo-controlled trial is now possible with a more definitive set of outcomes that are meaningful for patients.

We don't see any easy answers here—long-term outcomes would require a long-term study, after all, and clinical decisions need to be made right away, every day—but perhaps we can use our examination of this particular study and its reporting to suggest practical directions for improvement in heart treatment studies and more generally.

References

American College of Cardiology (2017). ORBITA: First placebo-controlled randomized trial of PCI in CAD patients. *ACC News*, 2 Nov. http://www.acc.org/latest-in-cardiology/articles/2017/10/27/13/34/thurs-1150am-orbita-tct-2017

Al-Lamee, R., Thompson, D., Dehbi, H. M., Sen, S., Tang, K., Davies, J., Keeble, T., Mielewczik, M., Kaprielian, R., Malik, I. S., Nijjer, S. S., Petraco, R., Cook, C., Ahmad, Y., Howard, J., Baker, C., Sharp, A., Gerber, R., Talwar, S., Assomull, R., Mayet, J., Wensel, R., Collier, D., Shun-Shin, M., Thom, S. A., Davies, J. E., and Francis, D. P. (2017). Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial. *Lancet*. http://dx.doi.org/10.1016/S0140-6736(17)32714-9

Belluz, J. (2017). Thousands of heart patients get stents that may do more harm than good. Vox.com, 6 Nov. https://www.vox.com/science-and-health/2017/11/3/16599072/stent-chest-pain-treatment-angina-not-effective

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Harrell, F. (2017a). Statistical errors in the medical literature. Statistical Thinking blog, 8 Apr. http://www.fharrell.com/2017/04/statistical-errors-in-medical-literature.html

Harrell, F. (2017b). Statistical criticism is easy; I need to remember that real people are involved. Statistical Thinking blog, 5 Nov. http://www.fharrell.com/2017/11/statistiorbita-tct-2017cal-criticism-is-easy-i-need-to.html

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, 16–23.

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., and Stern, H. S. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician* **60**, 328–331.

Kolata, G. (2017). 'Unbelievable': Heart stents fail to ease chest pain. *New York Times*, 2 Nov. https://www.nytimes.com/2017/11/02/health/heart-disease-stents.html

O'Hare, R. (2017). Study of heart stents for stable angina highlights potential of placebo effect. Imperial College London, 2 Nov. http://www3.imperial.ac.uk/newsandeventspggrp/

imperialcollege/newssummary/news_2-11-2017-15-52-46

Resnick, B. (2017). White fear of demographic change is a powerful psychological force. Vox.com, 28 Jan. https://www.vox.com/science-and-health/2017/1/26/14340542/white-fear-trump-psychology-minority-majority

Sands, M. L. (2017). Exposure to inequality affects support for redistribution. *Proceedings of the National Academy of Sciences* **114**, 663–668.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359-1366.

Vickers, A. J., and Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *British Medical Journal* **323**, 1123–1124.