



Objective Randomised Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina (ORBITA) and coronary stents: A case study in the analysis and reporting of clinical trials

Andrew Gelman,^a John B. Carlin,^b and Brahmajee K. Nallamothu,^c *New York City, NY; Melbourne, Australia; and Ann Arbor, MI*

In the fall of 2017, Al-Lamee et al¹ reported results from a randomized controlled trial of percutaneous coronary intervention using coronary stents for stable angina. The study, called ORBITA, included approximately 200 patients and 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 performed. In follow-up, patients were asked to guess their treatment, and of those who were willing to guess, only 56% guessed correctly, indicating that the blinding was largely successful.

The summary finding from the study was that stenting did not “increase exercise time by more than the effect of a placebo procedure,” with the mean difference in this primary outcome between treatment and control groups reported as 16.6 seconds with an SE of 9.8 (95% CI, -8.9 to +42.0 seconds) and a *P* value of .20. In the *New York Times*, Kolata² reported the finding as “unbelievable,” remarking that it “stunned leading cardiologists by countering decades of clinical experience.” Indeed, one of us (B.K.N.) was quoted as being humbled by the finding, as many cardiologists had expected a positive result. On the other hand, Kolata noted, “there have long been questions about [stents'] effectiveness.” At the very least, the willingness of doctors and patients to participate in a controlled trial with a placebo procedure suggests some degree of existing skepticism and clinical equipoise.

ORBITA was a landmark trial due to its innovative use of blinding for a surgical procedure.¹⁹ However, substantial questions remain regarding the role of stenting in stable angina. It is a well-known statistical fallacy to take a result that is not statistically significant and report it as zero, as was essentially done here based on the *P* value of .20 for the primary outcome. Had this comparison happened to produce a *P* value of .04, would the headline have been, “Confirmed: Heart Stents Indeed Ease Chest Pain”? A lot of certainty seems to be hanging on a small bit of data.

The purpose of this article is to take a closer look at the lack of statistical significance in ORBITA and the larger questions this trial raises about statistical analyses, statistically based decision making, and the reporting of clinical trials. This review of ORBITA is particularly timely in the context of the widely publicized statement released by the American Statistical Association that cautioned against the use of sharp thresholds for the interpretation of *P* values³ and more recent extensions of this advice by ourselves and others.^{4,5} We end by offering potential recommendations to improve reporting.

Statistical analysis of the ORBITA trial Adjusting for baseline differences

In ORBITA, exercise time in a standardized treadmill test—the primary outcome in the preregistered design—increased on average by 28.4 seconds in the treatment group compared with an increase of only 11.8 seconds in the control group. As noted above, this difference was not statistically significant at a significance threshold of .05. Following conventional rules of scientific reporting, the true effect was treated as zero, an instance of the regrettably common statistical fallacy of presenting nonstatistically significant results as confirmation of the null hypothesis of no difference.

However, the estimate using gain in exercise time does not make full use of the data that were available on differences between the comparison groups at

From the ^aDepartment of Statistics and Political Science, Columbia University, New York City, NY, ^bClinical Epidemiology and Biostatistics, Murdoch Children's Research Institute, Melbourne School of Population and Global Health and Department of Paediatrics, University of Melbourne, Melbourne, Australia, and ^cMichigan Integrated Center for Health Analytics and Medical Prediction, Division of Cardiovascular Diseases, Department of Internal Medicine, University of Michigan Medical School, Ann, MI.

Submitted April 1, 2019; accepted April 2, 2019.

Reprint requests: Brahmajee K. Nallamothu, Department of Internal Medicine, University of Michigan Medical School, Building 16, NCRC Rm 132W, Ann Arbor, MI 48109.

E-mail: bnallamo@med.umich.edu

0002-8703

Published by Elsevier Inc.

<https://doi.org/10.1016/j.ahj.2019.04.011>

Table I. Summary data comparing stents to placebo, from Table 3 of Al-Lamee et al¹

Measurement	n	Treatment			n	Control			Comparison	
		Pre \bar{y} (SD)	Post \bar{y} (SD)	Gain diff (CI)		Pre \bar{y} (SD)	Post \bar{y} (SD)	Gain diff (CI)	Est (CI)	<i>p</i>
Exercise time (s)	104	528.0 (178.7)	556.3 (178.7)	28.4 (11.6 to 45.1)	90	490.0 (195.0)	501.8 (190.9)	11.8 (-7.8 to 31.3)	16.6 (-8.9 to 42.0)	.200
Peak oxygen uptake (mL/min)	99	1715.0 (638.1)	1713.0 (583.7)	-2.0 (-54.1 to 50.1)	89	1707.4 (567.0)	1718.3 (550.4)	10.9 (-47.2 to 69.0)	-12.9 (-90.2 to 64.3)	.741
SAQ, physical limitation	100	71.3 (22.5)	78.6 (24.0)	7.4 (3.5 to 11.3)	88	69.1 (24.7)	74.1 (24.7)	5.0 (0.5 to 9.5)	2.4 (-3.5 to 8.3)	.420
SAQ, angina frequency	103	79.0 (25.5)	93.0 (26.8)	14.0 (9.0 to 18.9)	90	75.0 (31.4)	84.6 (27.7)	9.6 (3.6 to 15.5)	4.4	.260
SAQ, angina stability	102	64.7 (25.5)	60.5 (23.7)	-4.2 (-10.7 to 2.4)	89	68.5 (24.3)	63.5 (25.6)	-5.1 (-11.7 to 1.6)	0.9 (-8.4 to 10.2)	.851
EQ-5D-5 L QOL	103	0.80 (0.21)	0.83 (0.21)	0.03 (0.00 to 0.06)	89	0.79 (0.22)	0.82 (0.20)	0.03 (0.00 to 0.07)	0.00 (-0.04 to 0.04)	.994
Peak stress wall motion index score	80	1.11 (0.18)	1.03 (0.06)	-0.08 (-0.11 to -0.04)	57	1.11 (0.18)	1.13 (0.19)	0.02 (0.03 to 0.06)	-0.09 (-0.15 to -0.04)	.0011
Duke treadmill score	104	4.24 (4.82)	5.46 (4.79)	1.22 (0.37 to 2.07)	90	4.18 (4.65)	4.28 (4.98)	0.10 (-0.99 to 1.19)	1.12 (-0.23 to 2.47)	.104

Abbreviations: EQ, EuroQol; Est, estimate; QOL, quality-of-life; diff, difference; SAQ, Seattle Angina Questionnaire.

baseline.^{6,7} As can be seen in Table I, the treatment and placebo groups differ in their pretreatment levels of exercise time, with mean values of 528.0 and 490.0 seconds, respectively. This sort of difference is no surprise—randomization assures balance only in expectation—but it is important to adjust for this discrepancy in estimating the treatment effect. In a published article, the adjustment was performed by simple subtraction of the pretreatment values:

$$\text{Gain score estimated effect} : (y_{\text{post}} - y_{\text{pre}})^T - (y_{\text{post}} - y_{\text{pre}})^C, \quad (1)$$

However, this *overcorrects* for differences in pretest scores because of the familiar phenomenon of “regression to the mean”; just from natural variation, we would expect patients with lower scores at baseline to improve, relative to the average, and patients with higher scores to regress downward. The optimal linear estimate of the treatment effect is actually:

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

where β is the coefficient of y_{pre} in a least-squares regression of y_{post} on y_{pre} , also controlling for the treatment indicator. The estimate in (1) is a special case of the regression estimate (2) corresponding to $\beta = 1$. Given that the pretest and posttest measurements are positively correlated and have nearly identical variances (as can be seen in Table D), we can anticipate that the optimal β will be less than 1, which will reduce the correction for difference in pretest and thus increase the

estimated treatment effect while also decreasing the SE. As a result, an adjusted analysis of these data would be expected to produce a lower *P* value.

The adjusted regression analysis can be done using the information available in Table I, as explained in detail in Box 1. The *P* value from this adjusted analysis is .09: as anticipated, lower than the *P* = .20 from the unadjusted analysis.

Alternative reporting

Despite moving closer to the conventional .05 threshold, the *P* value of .09 remains above the traditional level of significance at which one is taught to reject the null hypothesis. A potential blockbuster reversal with an adjusted analysis—“Statistical Sleuths Turn Reported Null Effect into a Statistically Significant Effect”—does not quite materialize.

However, within different conventions for scientific reporting, this experiment could have been presented as positive evidence in favor of stents. In some settings, a *P* value of .09 is considered statistically significant; for example, in a recent social science experiment published in the *Proceedings of the National Academy of Sciences*, Sands⁹ presented a causal effect based on a *P* value of less than .10, and this was enough for publication in a top journal and in the popular press, with, for example, that work mentioned uncritically in the media outlet *Vox* without any concern regarding significance levels.¹⁰ By contrast, *Vox* reported that ORBITA showed stents to be “dubious treatments,” a prime example of the “epidemic of unnecessary medical treatments”.¹¹ Had Al-Lamee et al performed the adjusted analysis with their data and

Box 1

Using the reported data summaries to obtain the analysis controlling for the pretreatment measure.

For each of the treatment and control groups, we are given the SD of the pretest measurements, the SD of the posttest measurements, and the SD of their difference, which can be obtained by taking the width of the CI for the difference, dividing by 4 to get the SE of the difference, and then multiplying by \sqrt{n} to get back to the SD.

Then we use the rule, $sd(y_2 - y_1) = \sqrt{sd(y_1)^2 + sd(y_2)^2 - 2sd(y_1)sd(y_2)}$ 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, $\beta = \rho sd(y_2)/sd(y_1)$, which yields $\beta = 0.88$ for the treated and $\beta = 0.86$ for the control group. If these 2 coefficients were much different from each other, we might want to consider an interaction model,⁸ but here they are close enough that we simply take their average.

We use the average, $\beta = 0.87$, in (2) and get an estimate for the adjusted mean difference of 21.3 (indeed, quite a bit higher than the reported difference in gain scores of 16.6) with an SE of 12.5 (very slightly lower than 12.7, the SE of the difference in gain scores) and 95% CI -3.2 to 45.8 s. The estimate is not quite 2 SEs away from zero: the z score is 1.7, and the P value is .09.

published in *PNAS* rather than the *Lancet*, could they have confidently reported a causal effect of stents on exercise time?

Our point here is not at all to suggest that Al-Lamee et al engaged in reverse “p-hacking,”¹² choosing an analysis that produced a newsworthy null result. In fact, the authors should be congratulated for preregistering their study, publishing their protocol before performing their analyses, and reporting a prespecified primary analysis. Rather we wish to emphasize the flexibility inherent both in data analysis and reporting, even in the case of a clean randomized experiment. We are pointing out the potential fragility of the ‘stents-didn't-work story’ in this case. Existing data could easily have been presented as a success for stents compared with placebo by authors who were aiming for that narrative and performing reasonable analyses.

Fragility of the findings

How sensitive were the results to slight changes in the data? To better understand this critical point, one can perform a simple bootstrap analysis, computing the results that would have been obtained from reanalyzing the data 1,000 times, each time resampling patients from the existing experiment with replacement.¹³ Because raw data were not available to us, 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 traditional level of statistical significance. This is not to say that stents really are better than placebo; the data also seem consistent with a null effect. The take-home point of this experiment is that the results could easily have gone “the other way,” when reporting is forced into a binary classification of statistical significance, for many different reasons.

Design of the trial and clinical significance

In a justification for their study design and sample size, Al-Lamee et al¹ wrote: “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 was conservatively designed to be able to detect an effect size of 30 s.” The estimated effect of 21 seconds with SE 12 seconds is consistent with the “conservative” effect size estimate of 30 seconds given in the published article. So, although the experimental results are consistent with a null effect, they are even more consistent with a small positive effect.

One might ask, however, about the *clinical* significance of such a treatment effect, which we can discuss without relevance to P values or statistical significance. For simplicity, 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? One way to conceptualize this is in terms of percentiles. The data show a prerandomization distribution (averaging the treatment and control groups) with a mean of 509 and a SD 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 a real population shift.

Beyond exercise time, there were other signals from ORBITA that seemed to suggest consistent improvements in the physiological parameter of ischemia through end points such as fractional flow reserve, instantaneous wave-free ratio, and stress echo. Actually, findings from the stress echo highlight a potentially important avenue into an alternative presentation of these results. There is no question that some physiological changes are being

made by stents, with very large and highly statistically significant ($P < .001$) effects seen on echo measures. As is often the case, the null hypothesis that these physical changes should make absolutely zero difference to any downstream clinical outcomes seems farfetched. Thus, the sensible question to ask is “How large are the clinical differences observed?” not “How surprising is the observed mean difference under a [spurious] null hypothesis?” The simple textbook way to tackle this question is to report CIs around the mean differences and not to focus on whether the intervals happen to include zero. That the standard 95% CI for the primary outcome comfortably includes the target effect size of 30 seconds suggests that this value should be no more “rejected” than the null value. Furthermore, without the longitudinal data to observe the outcomes that matter most to patients—health and length of life—much remains uncertain.

The larger question has to be about balancing the long-term benefits of stents with risks 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 on a standardized treadmill test, or even a hypothesized larger benefit of 50 seconds, which would still only represent a 10% improvement for an average patient in this study. However, maybe a 5% to 10% increase is consequential in this case, as it could improve quality of life for a patient outside this artificial setting. Perhaps this small gain in exercise time is associated with the need for fewer medications, fewer functional limitations, or greater mobility. If so, however, one might postulate that this gain would have been apparent in assessments of angina burden, and it was not.

Part of the bigger concern here is that these patients were already doing pretty well on medications; that is, they already had a low symptom frequency before stenting. For example, angina frequency as measured by the Seattle Angina Questionnaire was 63.2 after optimizing medications and before stenting in the treatment group. This roughly translates as “monthly” angina (John Spertus, personal communication). How does a study with a follow-up of just 6 weeks expect to improve an outcome that happens this infrequently? In fact, one of the great debates surrounding ORBITA is that those who discount the trial suggest it enrolled patients who typically do not receive stents in routine practice. Those who believe that ORBITA is a game changer argue that these less symptomatic patients actually make up a large proportion of those receiving stents, and that is why we have such a large problem with their overuse.

Finally, are stents really being given to patients with stable angina just to improve fitness or to reduce symptoms? Or is there a continued expectation that stents have long-term benefits for patients, despite earlier data from studies like the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation

(COURAGE) study?¹⁴ This would seem to be the key question, in which case the short-term effects, or lack thereof, found in the ORBITA study are largely irrelevant. Other larger trials, such as International Study of Comparative Health Effectiveness With Medical and Invasive Approaches (ISCHEMIA, see: <https://clinicaltrials.gov/ct2/show/NCT01471522>) are considering this more fundamental question but will not have a placebo procedure.

Recommendations for statistical reporting of trials

The search for better medical care is an incremental process, with incomplete evidence accumulating over time. There is unfortunately a fundamental incompatibility between that core idea and the common practice, both in medical journals and in the news media, of up-or-down reporting of individual studies based on statistical significance. We offer some recommendations summarized in **Box 2** that we believe will be helpful to authors and editors moving forward.

At this point, it is not clear how best to incorporate this recent experiment into routine practice despite its novel and provocative study design, so the forced reporting of the primary outcome as “positive” or “negative” is unhelpful. A reanalysis of the summary data from Al-Lamee et al¹ reveals a stronger estimated effect that is closer to the conventional boundary of statistical significance, indicating that the study could rather easily have generated and reported evidence in favor of, rather than against, the effectiveness of stents for patients with stable angina. 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 statistical analysis. In any case, though, the observed increases in exercise time, even if statistically significant, do not seem at first glance to be of much clinical importance compared with the much more relevant long-term health outcomes that remain uncertain.

In the design, evaluation, and reporting of experimental studies, there is a norm of focusing on the statistical significance of a primary outcome: in this case, change in average exercise time on a standardized treadmill test. In general, the conclusions that follow will be fragile because P values are extremely noisy unless the underlying effect is huge. An experiment may be designed to have 80% power, but this does not eliminate the fragility, as illustrated by our bootstrap re-analysis. Power calculations are often conditional on an over-estimated effect size^{15,16} and do not address the important question of variation in treatment effects. Examination of the *Lancet* article and its reception in the news media suggests that it exhibits a classic case of “significantitis” or “dichotomania,”¹⁷ with frequent

Box 2 Recommendations for analyses and reporting.

Analyses

1. Baseline adjustment for differences: should be prespecified for the primary analysis where strong confounders such as a baseline measure of the outcome are available.
2. Be aware of fragility of inferences. Fragility can be demonstrated using the sampling or posterior distribution as estimated using mathematical modeling, bootstrap simulation, or Bayesian analysis.

Reporting

1. Avoid use of sharp thresholds for *P* values and thus eliminate the term “statistical significance” from the reporting of results.
2. Consider the full range (upper and lower ends) of interval estimates for important outcomes and their potential inclusion of clinically important differences.
3. Consider the potential for individual variability in responses (heterogeneity of treatment effects) and not just mean differences.

repetition of phrases such as “there was no significant difference.” In line with current thinking,⁴ we suggest that the phrase used by these authors, “We deemed a *p* value less than 0.05 to be significant,” should be strongly discouraged, rather than actively demanded as is currently the case by many journal editors. To their credit, the ORBITA authors themselves have recognized these critical issues (see <https://twitter.com/ProfDFrancis/status/952008644018753536>).

In the case of stents, an important disconnect appears between the findings emphasized in the recent study—however presented—and the larger context of treatments for heart disease. From a statistical perspective, this seems to reflect a problem with the framing of clinical trials as attempts to discover whether a treatment has a statistically significant effect, commonly misinterpreted to be equivalent to a real (nonzero) population mean difference. Power calculations are used in an attempt to assure stable estimates and a good chance of the experiment being “successful,” although within these constraints, there can be a push toward convenience rather than relevance of outcome measures²⁰, which is perhaps an inevitable compromise. ORBITA shows us the confusion that arises when a treatment is reported as a success or failure in statistical terms that are divorced from clinical context.

ORBITA was never meant to be definitive in a broad sense; it was designed to find a statistically significant physiological effect of stenting on mean exercise time, without clarity on the clinical relevance of anticipated effects on this outcome measure. Indeed, a likely 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 in this setting. Further background on these topics from Darrel Francis, the senior author on the study, appears at Harrell.¹⁸ Beyond immediate news reports, one positive impact of ORBITA is that bigger trials of stenting with placebo procedures are now much more likely with a more definitive set of measured outcomes that are meaningful for patients.

We do not 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 in the design and reporting of clinical trials more generally.

Acknowledgments

We thank Doug Helmreich for bringing this example to our attention, Shira Mitchell for helpful comments, and the Office of Naval Research, Defense Advanced Research Project Agency, and the National Institutes of Health for partial support of this work.

Disclosures

Dr. Gelman and Dr. Carlin report no competing interests. Dr. Nallamothu is an interventional cardiologist and Editor-in-Chief of a journal of the American Heart Association but otherwise has no competing interests.

References

1. Al-Lamee R, Thompson D, Dehbi HM, et al. Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial. *Lancet* 2017 [https://doi.org/10.1016/S0140-6736\(17\)32714-9](https://doi.org/10.1016/S0140-6736(17)32714-9).
2. Kolata G. ‘Unbelievable’: heart stents fail to ease chest pain. *New York Times* 2017.2 Nov. <https://www.nytimes.com/2017/11/02/health/heart-disease-stents.html>.
3. Wasserstein RL, Lazar NA. The ASA’s statement on *p*-values: context, process, and purpose. *American Statistician* 2016;70:129-33.
4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305-7.
5. McShane BB, Gal D, Gelman A, et al. Abandon statistical significance. *American Statistician* 2019;73(S1):235-45.
6. Harrell F. Statistical errors in the medical literature. *Statistical Thinking blog* 2017(8 Apr) <http://www.fharrell.com/2017/04/statistical-errors-in-medical-literature.html>.

7. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. *Br Med J* 2001;323:1123-4.
8. Gelman A. Treatment effects in before-after data. In: Gelman A, Meng XL, eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: Wiley; 2004.. Chapter 18.
9. Sands ML. Exposure to inequality affects support for redistribution. *Proc Natl Acad Sci* 2017;114:663-8.
10. 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>
11. 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>
12. Simmons J, Nelson L, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychol Sci* 2011;22:1359-66.
13. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979;7:1-26.
14. Boden WE, O'Rourke RA, Teo KK, et al. Optimal medical therapy with or without PCI for stable coronary disease. *New England Journal of Medicine* 2007;356:1503-16. Epub 2007 Mar 26.
15. Gelman A. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Pers Soc Psychol Bull* 2018;44:16-23.
16. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;365:1348-53.
17. Greenland S. The need for cognitive science in methodology. *Am J Epidemiol* 2017;186:639-45.
18. 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/statistic-orbita-tct-2017-cal-criticism-is-easy-i-need-to.html>
19. 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>.
20. Gelman A, Carlin JB. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 2014;9:641-51.