ORBITA: A case study in the analysis and reporting of clinical trials

Andrew Gelman, John Carlin and Brahmajee K Nallamothu

14 Mar 2018

Department of Statistics and Political Science, Columbia University, New York City, NY, United States (Andrew Gelman, professor); Clinical Epidemiology & Biostatistics, Murdoch Children's Research Institute, Melbourne School of Population and Global Health and Department of Paediatrics, University of Melbourne, Melbourne, Australia (John Carlin, professor); Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, United States (Brahmajee K Nallamothu, professor);

Correspondence to: Brahmajee K Nallamothu bnallamo@med.umich.edu

Acknowledgements: 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.

Competing interests: 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.

Word Count: 2085

Introduction

ORBITA (Objective Randomised Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina) was a randomized clinical trial of approximately 200 patients in which half the patients received stents and half received a placebo procedure. Its summary finding 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 sec (95% confidence interval, –8.9 to +42.0 sec) and a p-value of 0.20.

In the *New York Times*, Kolata (2017) reported the finding as "unbelievable," remarking that it "stunned leading cardiologists by countering decades of clinical experience." Indeed, one of us (BKN) was quoted as being humbled by the finding as many 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 a placebo procedure. However, substantial questions remain even after ORBITA 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 0.20 for the primary outcome. Had this comparison happened to produce a p-value of 0.04, would the headline have been, "'Believable': Heart Stents Indeed Ease Chest Pain"?

The purpose of this paper is to take a closer look at the lack of statistical significance in ORBITA and the larger questions it raises about statistical analyses, statistically based versus clinical decision-making, and the reporting of clinical trials. This is important because a lot of certainty seems to be hanging on a small bit of data.

Dichotomized thresholds are a big problem, hence in this paper we will avoid discussing "statistical significance" except when discussing issues of how results are or could be reported.

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 sec in the treatment group compared to an increase of only 11.8 sec in the control group. As noted above, this difference was associated with a p-value greater than 0.05. Hence, following conventional rules of scientific reporting it was treated as zero—an instance of the regrettably common statistical fallacy of presenting non-statistically-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 groups at baseline (Vickers and Altman, 2001, Harrell, 2017a). The treatment and placebo groups differed in their pre-treatment levels of exercise time, with mean values of 528.0 and 490.0 s, respectively (**Supplementary Table**). This sort of difference is fine—randomization assures balance only in expectation—but it is important to adjust for this discrepancy in estimating the treatment effect. In the published paper, the adjustment was performed by simple subtraction of the pre-treatment values:

Gain in exercise time: $$(y_{post} - y_{pre})^{T} - (y_{post} - y_{pre})^{C}, \qquad (1)$$

But this *over-corrects* for differences in pre-test 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:

Gain in exercise time: $$(y_{post} - \beta\, y_{pre})^{T} - (y_{post} - \beta\, y_{pre})^{C}, \qquad (2)$$

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 pre-test and post-test measurements have nearly identical variances, 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 while decreasing the standard error.

An adjusted analysis using the information available is explained in detail in **Box 1.** The p-value from this adjusted analysis is 0.09: as expected, lower than the p=0.20 from the unadjusted analysis. What is relevant is not whether or not this new p-value has become

"statistically significant" but rather the reported p-value is subject to change based on alternative analyses.

Within different conventions for scientific reporting and for different fields, a p-value of 0.09 is considered to be 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 a p-value of less than 0.10, and this was enough for publication in a top journal and in the popular press. *Vox* mentioned that work uncritically without any concern regarding significance levels (Resnick, 2017). By contrast, *Vox* reported stents as a prime example of the "epidemic of unnecessary medical treatments" after ORBITA (Belluz, 2017).

These concerns are deepened further when one considers how sensitive results from ORBITA were from a statistical standpoint. To better understand this one can perform 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). As 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 with p-values less than 0.05. This is not to say that stents really are better on average than placebo in improving exercise time—the data also appear 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.

Statistically Based versus Clinical Decision-Making

In justifying their study design and sample size, Al-Lamee et al. (2017) wrote: "Evidence from placebo-controlled randomised controlled trials shows that single antianginal therapies provide improvements in exercise time of 48–55 sec…Given the previous evidence, ORBITA was conservatively designed to be able to detect an effect size of 30 sec." The estimated effect of 21 sec with standard error 12 sec is consistent with the "conservative" effect size estimate of 30 sec 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 sec? 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 sec and a standard deviation of 188 sec. Assuming a normal approximation, an increase in exercise time of 21 sec from 509 to 530 sec 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. Indeed, a recent study after ORBITA suggested ironically that such gains are possible during a treadmill test by simply playing music.

Thus, the larger clinical question is how to balance the long-term benefits of stents with risks of the procedure. It does not seem reasonable for a person to receive stents just for a potential benefit of 21 sec of exercise time on a standardized treadmill test—or even a hypothesized larger benefit of 50 sec, which would still only represent a 10% improvement for an average patient in this study. Yet maybe a 5% to 10% increase is consequential in this case as it could improve quality of life for a patient. Perhaps this small gain in exercise time is associated with the need for less medications, fewer functional limitations or greater mobility. If so, however, one might postulate this gain would have been apparent in assessments of angina burden, and it was not.

A big 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 ORBITA is a game-changer argue that these less symptomatic patients actually make up a large proportion of those receiving stents.

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 (Boden, 2007)? 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.

Evidence from ORBITA that pointed toward consistent improvements in the physiological parameter of ischemia through endpoints such as fractional flow reserve and stress echo suggests there is little question that some physiological changes are being made by stents, with very large and highly statistically significant. 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 and are they worth it?"—not "How surprising is the observed mean difference under a [spurious] null hypothesis?"

4. 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 the news media, of up-or-down reporting of individual studies based on statistical significance. We offer some recommendations to tackle this issue in **Box 2**.

In the design, evaluation, and reporting of experimental studies, there is a norm of focusing on the statistical significance of a primary outcome—described at times as "significantitis" or "dichotomania" (Greenland, 2017). It leads to an overreliance on phrases like, "We deemed a p value less than 0.05 to be significant," that are common throughout the published literature. The resulting conclusions from such a process frequently will be fragile because p-values are extremely noisy unless the underlying effect is huge. To their credit, the ORBITA authors themselves have recognized these critical issues (see online: https://twitter.com/ProfDFrancis/status/952008644018753536).

ORBITA was never meant to be definitive in a broad sense—it was designed to find a physiological effect of stenting on mean exercise time, without clarity on the clinical relevance of this outcome. Indeed, a likely reason why the study was limited to this endpoint 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 (2017b). One certain impact of ORBITA is that bigger trials of stenting with placebo procedures are now much more likely with a more meaningful set of outcomes that will be measured.

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 in the design and reporting of clinical trials more generally.

References

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

Allison, D. B., Brown, A. W., George, B. J., Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. Nature 530, 27–29. doi: 10.1038/530027a. PubMed PMID: 26842041; PubMed Central PMCID: PMC4831566.

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

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

Bland, J. M., and Altman, D. G. (2015). Best (but oft forgotten) practices: Testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. American Journal of Clinical Nutrition 102, 991–994. doi: 10.3945/ajcn.115.119768. Epub 2015 Sep 9. PubMed PMID: 26354536.

Boden, W. E., O'Rourke, R. A., Teo, K. K., Hartigan, P. M., Maron, D. J., Kostuk, W. J., Knudtson, M., Dada, M., Casperson, P., Harris, C. L., Chaitman, B. R., Shaw, L., Gosselin, G., Nawaz, S., Title, L. M., Gau, G., Blaustein, A. S., Booth, D. C., Bates, E. R., Spertus, J. A., Berman, D. S., Mancini, G. B., and Weintraub, W. S.; COURAGE Trial Research Group. (2007). Optimal medical therapy with or without PCI for stable coronary disease. New England Journal of Medicine 356, 1503–16. Epub 2007 Mar 26.

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

Gelman, A. (2004). Treatment effects in before-after data. In Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives, ed. A. Gelman and X. L. Meng, chapter 18. New York: Wiley.

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 Stern, H. S. (2006). The difference between “significant” and “not significant” is not itself statistically significant. American Statistician 60, 328–331.

Greenland, S. (2017). The need for cognitive science in methodology. American Journal of Epidemiology 186, 639–645.

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

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

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.

Schulz, K. F., and Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. Lancet 365, 1348–1353.

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.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. American Statistician 70, 129–133.

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

Measurement	Treatment				Control				Comparison	
	N	Pre $\bar{y}$ (sd)	Post $\bar{y}$ (sd)	Gain diff (ci)	N	Pre $\bar{y}$ (sd)	Post $\bar{y}$ (sd)	Gain 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

Box 1. Using the reported data summaries to obtain the analysis controlling for the pre-treatment measure

For each of the treatment and control groups, 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 get back to the standard deviation.

Then we use the rule, $\mathrm{sd}(\mathrm{y}_2 - \mathrm{y}_1) = \sqrt{\mathrm{sd}(\mathrm{y}_1)^2 + \mathrm{sd}(\mathrm{y}_2)^2 - 2\rho\,\mathrm{sd}(\mathrm{y}_1)\mathrm{sd}(\mathrm{y}_2)}$ and solve for ρ, the correlation between before and after measurements within each group. The result in this case is ρ = 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, $\beta = \rho\,\mathrm{sd}(\mathrm{y}_2)/\mathrm{sd}(\mathrm{y}_1)$, which yields $\beta = 0.88$ for the treated and $\beta = 0.86$ for the control group. If these two coefficients were much different from each other, we might want to consider an interaction model (Gelman, 2004), but here they are close enough that we simply take their average.

We use the average, β = 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 a standard error of 12.5 (very slightly lower than 12.7, the standard error of the difference in gain scores) and 95% CI –3.2 to 45.8 s. The estimate is not quite two standard errors away from zero: the z-score is 1.7, and the p-value is 0.09.

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.