Published in final edited form as:
J Am Coll Surg. 2013 January ; 216(1): 158–166. doi:10.1016/j.jamcollsurg.2012.09.015.

Are Surgical Trials with Negative Results Being Interpreted Correctly?

Baruch A Brody, PhD, **Carol M Ashton, MD, MPH**, **Dandan Liu**, **Youxin Xiong**, **Xuan Yao, BA**, and **Nelda P Wray, MD, MPH**
Department of Philosophy, Rice University (Brody, Liu, Xiong, Yao); the Center for Ethics, Medicine, and Public Policy (Brody); and The Methodist Hospital Research Institute and Department of Surgery (Ashton, Wray); Houston, TX

Abstract

Background—Many published accounts of clinical trials report no differences between the treatment arms, while being underpowered to find differences. This study determined how the authors of these reports interpreted their findings.

Study Design—We examined 54 reports of surgical trials chosen randomly from a database of 110 influential trials conducted in 2008. Seven that reported having adequate statistical power ($\beta \geq 0.9$) were excluded from further analysis, as were the 32 that reported significant differences between the treatment arms. We examined the remaining 15 to see whether the authors interpreted their negative findings appropriately. Appropriate interpretations discussed the lack of power and/or called for larger studies.

Results—Three of the 7 trials that did not report an a priori power calculation offered inappropriate interpretations, as did 3 of the 8 trials that reported an a priori power < 0.90. However, we examined only a modest number of trial reports from 1 year.

Conclusions—Negative findings in underpowered trials were often interpreted as showing the equivalence of the treatment arms with no discussion of the issue of being underpowered. This may lead clinicians to accept new treatments that have not been validated.

Treatments are often recommended for clinical practice after trials find no significant differences between these treatments and the current standard of care. However, these trials may be conducted with insufficient power. Under conditions of insufficient power, one must be careful in the interpretation of the meaning of a finding of no difference between treatment arms; ineffective treatments might mistakenly be recommended because the data show no difference between the treatment arms. Can one conclude that there really is no

Correspondence address: Baruch A Brody, PhD, Department of Philosophy, Rice University, Houston, TX 77251. bbrody@rice.edu.

Disclosure Information: Nothing to disclose.

Author Contributions: Study conception and design: Brody, Ashton, Wray
Acquisition of data: Brody, Ashton, Liu, Xiong, Yao, Wray
Analysis and interpretation of data: Brody, Ashton, Liu, Xiong, Yao, Wray
Drafting of manuscript: Brody, Ashton, Liu, Xiong, Yao, Wray
Critical revision: Brody, Ashton, Liu, Xiong, Yao, Wray

difference in the treatments, so the recommendation of the new treatment is justified, or is the lack of a difference simply related to the small sample sizes, so the introduction of the new treatment is not justified?

The extent of this problem has not been adequately explored. An opportunity to study it arose as part of a larger project we are conducting to analyze the methodologic and ethical strengths and weaknesses of influential comparative surgical trials whose results were reported between 2000 and 2008.[1] In a sample of 290 surgical trials, we observed 130 that did not report a priori power calculations or left out important components of the power calculation. However, in many cases, the authors of these trials claimed no significant differences between the treatment arms and made clinical recommendations to introduce the treatment.

We therefore undertook this study to evaluate how the findings of "no significant differences between the treatment groups" were interpreted in trials that either did not report an adequate a priori statistical power (and may not have based their sample sizes on power considerations) or did report power calculations but with sample sizes that conferred low power ($\beta < 0.90$) to detect real differences between treatment groups.

Methods

This study is part of a project funded by the National Institutes of Health called "Ethical and Methodological Standards for Clinical Trials of Invasive Procedures."[2,3] The overall goal of the parent project is to develop and disseminate reasonable contemporary ethical and methodologic standards for trials of surgical and minimally invasive procedures. One step of the project was an analysis of the methodologic and ethical strengths and weaknesses of 290 influential comparative surgical trials whose results were reported between 2000 and 2008.[1]

Sample

The trials reviewed for this study comprised a 50% random sample of the 110 trials from the parent study that were published in 2008 (the most recent year available). Fifty-four were selected for review. For each trial, we reviewed only the publication in which the main results were reported.

Development of the review process

The first step was to systematize the processes of review and the data elements to be extracted from the trial reports, which would enable us to assess the actual trial results and the authors' interpretations of the findings of their trials. To achieve this, we first selected from the parent database of trial dossiers 5 trials that varied by important characteristics such as funding source, single or multiple performance sites, superiority or noninferiority design, type of control (alternative invasive procedure or noninvasive comparison group), allocation method, and methods of blinding. All 5 publications were then independently read by each of the 3 senior investigators and each of the 3 research assistants. Each reviewer outlined a process of review and noted those variables they believed important to evaluate findings and authors' interpretations. The 6 investigators then met to discuss their proposed review procedures and data elements. The following processes and data abstraction for the

articles were agreed on and used with the 54 trial reports. Each report would be read independently by 2 of the research assistants, who would record in spreadsheets the following data elements: publication title; journal; trial objectives as reported by the authors; trial intent (test of superiority or noninferiority); a description of the intervention and the control; a description of the study population studied, including the number of subjects in each treatment arm; the primary outcome (the endpoint designated by the author as the primary outcome, or, if a power calculation was reported, the outcome used in that calculation); the report of a power calculation for the primary outcome, and if present, the components of the power calculation that were reported; all secondary outcomes and any power calculations for these variables; the direction, magnitude and statistical significance of all findings; the authors' interpretation of the findings (abstracted from the abstract and/or the discussion section); and our assessment of the appropriateness of the authors' interpretation. Each of the 54 trials was also classified as to whether it met 1 or more of our categories of specific interest, namely, the absence of reported a priori statistical power calculations, or low statistical power (defined as $\beta < 0.090$).

Assessment of the appropriateness of authors' interpretations

There were no issues of interpretation in trials that showed a statistically significant difference between the 2 arms. Specific criteria were developed for assessing the appropriateness of authors' interpretation of trials finding no significant differences between treatment groups. First, for trials not reporting a priori power calculations or with low statistical power ($\beta < 0.90$), and finding “no significant differences” between treatment groups, the authors of the trial report had to include an explicit consideration of power in order for us to classify the interpretation as correct. We judged these reports to have an appropriate interpretation of the findings if the authors explicitly discussed the lack of power or small sample sizes as a potential explanation of the findings or if they recommended larger studies to confirm their findings. We judged authors' interpretations to be flawed when the authors interpreted “no significant difference” findings as indicating that the interventions in the study were equivalent or just as good as one another and failed to discuss statistical power or small sample sizes, or failed to recommend larger studies to confirm the findings.

Review process

To ensure consistency in data element recognition, each research assistant re-reviewed and abstracted the 5 test trial reports. These data were then presented and discussed at face-to-face meetings with the 3 senior investigators until all discrepancies were resolved and explained.

After training was completed, each of the 54 trial reports was independently reviewed and abstracted by 2 research assistants, who then met to discuss their findings. Any discrepancies were resolved or referred to the senior investigators to adjudicate.

To ensure the validity of the data, throughout the review process during face-to-face meetings of the entire team, 1 or more of the 3 senior investigators reviewed all abstracted data for clarity and internal consistency. Questions regarding the data were resolved by re-

review of the primary article. Finally, a senior investigator re-reviewed the classification of all 54 articles as to whether the articles did not report a priori power or had low power and claimed no significant difference and confirmed whether or not a misinterpretation was present.

Results

A high proportion of the 54 surgical trials (47 of 54) either did not report statistical power calculations or reported them and were underpowered. Of the 54 surgical trials, 23 (42.6%) did not report a priori power calculations. Twenty-four (44.4%) reported a power < 0.9; only 7 (13.0%) reported a power ≥ 0.9. Approximately one-third (15) of the reports that either did not include an a priori power calculation or had low power claimed no significant difference, and therefore met our criteria for further review (Fig. 1).

Of the 7 articles[4-10] that claimed no significant difference with unreported a priori power calculations, we judged the authors' interpretation regarding a claim of no significant difference as appropriate in 4 (Table 1). Three of these noted that small sample size and low power could be an explanation of the null finding and recommended that larger trial studies be undertaken to confirm the findings. One study only recommended a larger study to confirm their results. In 3 trials, however, the authors interpreted the finding of “no significant difference” as meaning the 2 treatments are equally good alternatives and did not consider low power as a potential explanation. Our post hoc calculations of the power of these studies showed that the power was very limited, even to detect substantial differences between the treatment groups.

The data on the 8 trials[11-18] that met our criteria of low a priori power (<0.9), although finding no significant difference between treatment groups, are provided in Table 2. Seven of these trials reported a power ≥ 0.8 and the other a power of 0.7. Four of these trials were judged to have appropriate interpretation of the null finding because authors discussed the limitation of power and suggested other studies (n = 3) or just suggested a larger study (n = 1). An additional trial was judged to have an appropriate interpretation because the investigators entered far more patients than the power calculation suggested would be needed, strengthening the claim of no significant difference. The authors of the 3 remaining trials did not consider low power as a potential explanation of their findings and did not suggest larger studies; they inappropriately interpreted their finding as meaning the 2 treatments are equally good alternatives.

Authors made inappropriate interpretations in 6 of the 15 (40%) trials we reviewed in which no significant differences were observed between treatment groups. They mistakenly concluded that the 2 treatments are equally good alternatives.

Discussion

Of the 54 trials we reviewed, only 7 (16%) reported adequate power to detect statistically significant differences between treatment groups. Of the 47 that did not, 15 (33%) reported no significant difference between the 2 groups. These are the studies of greatest concern because authors may make inappropriate clinical recommendations based on the findings. In

fact, 6 of these 15 (40%) drew the inappropriate conclusions that the 2 treatments being studied are equally good alternatives. This mistake may have resulted in clinicians adopting treatments with inadequate justification, potentially leading to adverse or suboptimal outcomes for their patients.

As early as 1978, Frieman and colleagues[19] called attention to the problem of "negative trials" with low power, which could have missed important differences between treatment arms. Since then, an extensive debate has raged over the ethics of conducting underpowered trials, with some[20] claiming that they are unethical except in certain special circumstances and others[21] offering a more positive evaluation of them as long as the reports of the published trials properly interpret the findings. Many studies[22,23] have reported on the prevalence of underpowered surgical trials as well as trials in other areas of medicine.[24] What makes our study different is that we report on how these underpowered trials were interpreted in the published reports. Our crucial finding is that the underpowered trials are often misinterpreted in the published reports of their findings. Others have offered preliminary data suggesting that these misinterpretations are often adopted even by those well-trained in statistics.[25]

The major implication of our study is that treatments are being recommended that may be inferior to the current standard of care. Clinicians have to take care to evaluate statistical power issues before accepting recommendations found in the literature about potential treatment options. The problem, however, lies with the trialists who, knowingly or unknowingly, misrepresent the meaning of underpowered trials and with the journal editors who publish flawed reports of trials. Patients may receive inadequate treatment unless clinicians, researchers, reviewers, and journal editors are more careful about the design and interpretation of underpowered trials.

Limitations

One major limitation of this study is that only surgical trials were reviewed, given the focus of our project. We have no way of evaluating whether the problem is greater or less in trials of nonsurgical interventions. Furthermore, because we studied only trials from 2008, we cannot determine how pervasive the issue was before 2008 and whether it has improved since then.

The most significant limitation to our study is the sample size. We studied only 54 surgical trials. Even though these were influential studies, which might be expected to be more methodologically sound than most trials, generalizing from these results to all surgical trials would be problematic. We have established the existence of these serious misinterpretations. Additional studies, with larger sample sizes, would be required to more firmly estimate the extent of this problem in the general surgical literature.

Also, we have called attention to the possibility that these misinterpretations may lead to inappropriate changes in actual practice. Our study was not designed to determine whether these changes actually took place. Further studies, involving billing or coding data sets from before and after the publication, would be required to show that these changes actually took place.

Conclusions

It is time to put an end to this problem. Journal editors and reviewers should pay more attention to authors' interpretations of underpowered studies, rejecting articles that contain such misrepresentations or requiring them to be revised. The clinical research community needs to better educate potential researchers about the importance of running adequately powered trials and about properly interpreting trials that are underpowered. Clinicians need to evaluate issues of power and interpretation before accepting recommendations emanating from clinical trials. All these measures are necessary if patients are to benefit from clinical trials and not be exposed to potential harm by being subjected to invasive procedures, which however promising, have not yet been shown to be superior, or at least not inferior, to the standard of care.

Acknowledgments

We wish to acknowledge the efforts of Danielle Wenner, PhD, currently at the Cleveland Clinic, for her efforts in compiling the database from which we chose the articles to review.

This work was funded by the National Institutes of Health grant # R01CA134995.

References

1. Wenner DM, Brody BA, Jarman AF, et al. Do surgical trials meet the ethical and scientific standards for clinical trials? J Am Coll Surg. 2012; 215:722–730. [PubMed: 22819638]
2. Ashton CM, Wray NP, Jarman AF, et al. Ethics and methods in surgical trials. J Med Ethics. 2009; 35:579–583. [PubMed: 19717699]
3. Ashton CM, Wray NP, Jarman AF, et al. A taxonomy of multinational ethical and methodological standards for clinical trials of therapeutic interventions. J Med Ethics. 2011; 37:368–373. [PubMed: 21429960]
4. Xu YD, Ou YK, Zheng YQ, et al. The treatment for postirradiation otitis media with effusion: a study of three methods. Laryngoscope. 2008; 118:2040–2043. [PubMed: 18818551]
5. Murphy DJ, Macleod M, Bahl R, et al. A randomized controlled trial of routine versus restrictive use of episiotomy at operative vaginal delivery: a multicenter pilot study. BJOG. 2008; 115:1695–1702. [PubMed: 19035944]
6. Meretoja TJ, von Smitten KA, Kuokkanen HO, et al. Complications of skin-sparing mastectomy followed by immediate breast reconstruction. Ann Plast Surg. 2008; 60:24–28. [PubMed: 18281791]
7. Lee KS, Choo MS, Lee YS, et al. Prospective comparison of the ‘inside-out’ and ‘outside-in’ transobturator-tape procedures for the treatment of female stress urinary incontinence. Int Urogynecol J Pelvic Floor Dysfunct. 2008; 19:577–584. [PubMed: 17940717]
8. Kempfert J, Opfermann UT, Richter M, et al. Twelve-month patency with the PAS-Port proximal connector device: a single center prospective randomized trial. Ann Thorac Surg. 2008; 85:1579–1584. [PubMed: 18442542]
9. Alvarez B, Ribo M, Maeso J, et al. Transcervical carotid stenting with flow reversal is safe in octogenarians: a preliminary safety study. J Vasc Surg. 2008; 47:96–100. [PubMed: 18060727]
10. Pearce C, Torres C, Stallings S, et al. Elective appendectomy at the time of cesarean delivery: a randomized controlled trial. AJOG. 2008; 199:491.
11. Glineur D, Hanet C, Poncelet A, et al. Comparison of bilateral internal thoracic artery revascularization using in situ or Y graft configurations. Circulation. 2008; 118(14 Suppl):S216–S221. [PubMed: 18824757]

12. Barber MD, Kleeman S, Karram MM, et al. Transobturator tape compared with tension-free vaginal tape for the treatment of stress urinary incontinence. Obstet Gynecol. 2008; 111:611–621. [PubMed: 18310363]
13. Rinne K, Laurikainen E, Kivelä A, et al. A randomized trial comparing TVT with TVT-O: 12-month results. Int Urogynecol J Pelvic Floor Dysfunct. 2008; 19:1049–1054. [PubMed: 18373046]
14. Kelbaek H, Terkelsen CJ, Helgvist S, et al. Randomized comparison of distal protection versus conventional treatment in primary percutaneous coronary intervention: the drug elution and distal protection in ST-elevation myocardial infarcation (DEDICATION) trial. J Am Coll Cardiol. 2008; 51:899–905. [PubMed: 18308157]
15. Jensen BO, Rasmussen LS, Steinbrüchel DA. Cognitive outcomes in elderly high-risk patients 1 year after off-pump versus on-pump coronary artery bypass grafting. A randomized trial. Eur J Cardiothorac Surg. 2008; 34:1016–1021. [PubMed: 18778948]
16. Ang KL, Chin D, Leyva F, et al. Randomized, controlled trial of intramuscular or intracoronary injection of autologous bone marrow cells into scarred myocardium during CABG versus CABG alone. Nat Clin Pract Cardiovasc Med. 2008; 5:663–670. [PubMed: 18711405]
17. Barbato JE, Dillavou E, Horowitz MB, et al. A randomized trial of carotid artery stenting with and without cerebral protection. J Vasc Surg. 2008; 47:760–765. [PubMed: 18295439]
18. Ferenc M, Gick M, Kienzle R, et al. Randomized trial on routine vs. provisional T-stenting in the treatment of de novo coronary bifurcation lesions. Eur Heart J. 2008; 29:2859–2867. [PubMed: 18845665]
19. Freiman J, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 “negative” trials. N Engl J Med. 1978; 299:690–694. [PubMed: 355881]
20. Halpern S, Karlawish J, Berlin J. The continuing unethical conduct of underpowered clinical trials. JAMA. 2002; 288:358–362. [PubMed: 12117401]
21. Schultz KF, Grimes DA. Sample size calculations in randomized trials. Lancet. 2005; 365:1348–1353. [PubMed: 15823387]
22. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials literature: equivalency or error? Arch Surg. 2001; 136:796–800. [PubMed: 11448393]
23. Maggard MA, O'Connell JB, Liu JH, et al. Sample size calculations in surgery: are they correctly performed? Surgery. 2003; 134:275–279. [PubMed: 12947329]
24. Bedard PL, Krzyzanowska MK, Pintilie M, Tannock I. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. J Clin Oncol. 2007; 25:3482–3487. [PubMed: 17687153]
25. Lecoutre MP, Poitevineau J. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. Int J Psychology. 2003; 38:37–45.

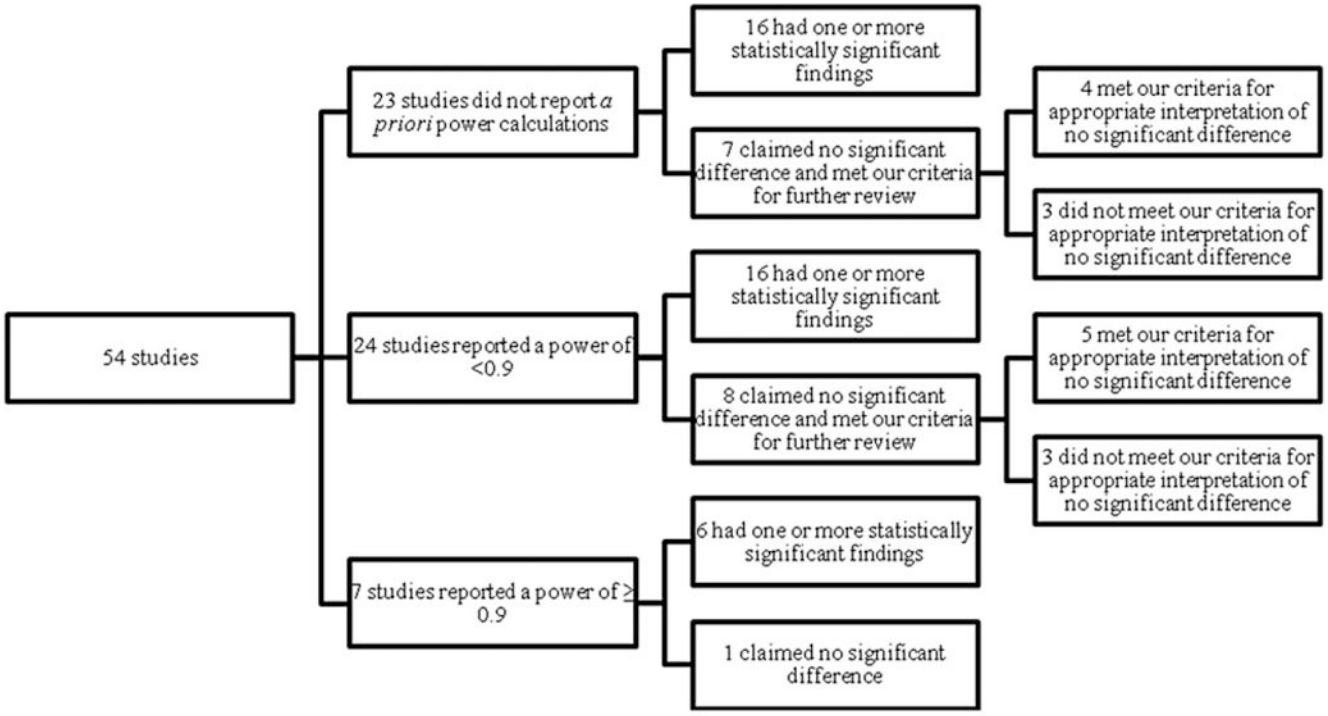

Figure 1.

Trials that met our criteria for further review to evaluate the appropriateness of the authors' interpretations of their findings.

Table 1

Characteristics of Clinical Trials that Did Not Report an A Priori Sample Size and Power Calculation and that Did Not Find a Significant Difference Between Treatment Groups

Study question	Primary endpoints	Secondary endpoints, n	Authors' conclusion	Our evaluation of the authors' interpretation
Evaluation of 3 treatments: simple auripuncture plus aspiration (n = 45); tympanic membrane fenestration with cauterization (n = 45); or myringotomy plus grommet insertion (n = 45) for postirradiation otitis media	Not designated	4	"The three methods used here each have advantages and disadvantages. In conclusion, we believe that a step by step approach should be adopted in choosing treatment methods. That is, we should use auripuncture first and then consider the other methods only if the former is inadequate."[4]	Interpretation inappropriate: power limitations were not discussed and large studies were not recommended. The authors claim equivalency even though all variables trend in favor of one intervention. Our calculations show the study had only a 50% power to detect a 37% increase in cure rate from 40% to 55%.
Restrictive use of episiotomy (n = 101) vs routine use of episiotomy (n = 99)	Extensive perineal tearing involving the anal sphincter (third- or fourth-degree tears)	19	"The pilot study does not provide conclusive evidence that a policy of routine episiotomy is better or worse than a restrictive policy. A definitive randomized controlled trial is feasible but will require a large sample size to inform clinical practice."[5]	Interpretation appropriate: power limitation of this "pilot study" discussed. A definitive randomized controlled trial was recommended and the sample size needed for that trial calculated.
High-frequency radiosurgery (n = 26) vs conventional diathermy (n = 38) to reduce complications after skin-saving mastectomy	Not designated	3	"This study shows that high-frequency radiosurgery is comparable to conventional diathermy in terms of complication rates. Further prospective randomized studies are required to critically evaluate the role of radiofrequency surgery and other newly developed dissection methods."[6]	Interpretation appropriate: though power limitations are not discussed, larger studies are recommended.
Comparison of the "inside-out" (n = 50) and "outside-in" (n = 50) transobturator-tape procedures for female stress urinary incontinence (SUI).	Not designated	22	"We would like to conclude that in our series, tension-free vaginal tape obturator (TVT-O) and transobturator-tape (TOT) appear equally effective for female SUI. However, this study was unable to identify a difference between the two procedures. The findings may be due to the underpowered nature of the study. Ideally, large well-constructed randomized controlled trial with longer follow-up period is necessary."[7]	Interpretation appropriate: power limitations were addressed and larger studies recommended.
PAS-Port (n = 51) vs conventional hand-sewn (n = 48) vein anastomosis to the aorta	Designated 2: patency at discharge and patency after 1 year	3	"This prospective randomized study demonstrated excellent short and midterm patency in both the hand-sewn and PAS-Port grafts. The PAS-Port system allowed for the rapid, safe, and effective creation of a proximal anastomosis without the need to clamp the aorta. Based on this study we consider this product a valid alternative for proximal anastomosis."[8]	Interpretation inappropriate: power limitations were not discussed and larger studies were not recommended. The authors claim equivalency even though all variables trend toward one intervention. Our calculations show that the study had only a 50% power to detect a

Study question	Primary endpoints	Secondary endpoints, n	Authors' conclusion	Our evaluation of the authors' interpretation
				50% reduction in complications: 25% to 12.5%
Transcervical carotid stenting with flow reversal (n = 36) vs carotid endarterectomy (n = 45)	Designated 3: stroke, death or myocardial infarct within 30 days	8	"The results of the intervention were comparable with the outcome of CEA in the same age group of patients. This study has the limitation of a small sample size, making the statistical power lower than required to derive definitive conclusion. Lastly, a long-term clinical follow-up is needed to guarantee the efficacy and longlasting benefits of this procedure."[9]	Interpretation appropriate: power limitations were discussed and larger studies were recommended.
Cesarean delivery and appendectomy (n = 45) vs standard cesarean delivery (n = 48)	Designated operative times and markers of morbidity.	2	"Our data suggest that appendectomy that is performed at the time of cesarean delivery does not increase inpatient maternal morbidity. Based on this, we believe that appendectomy at the time of cesarean delivery can be considered safely in selected patient."[10]	Interpretation inappropriate: though the study's limited power was presented in a post hoc analysis, it was not given consideration as a potential explanation for the null findings.

PAS, proximal anastomosis system

Table 2

Characteristics of Surgical Trials with an A Priori Power ≤ 0.90 to Detect a Difference and Findings of No Significant Difference Between Treatment Groups

Study question	Primary endpoints	Secondary endpoints, n	Authors' conclusion	Our evaluation of the authors' interpretation
Comparison of bilateral internal thoracic artery (BITA) revascularization using in situ (n = 152) or y graft (n = 152) configurations	Major adverse cerebro-cardiovascular events	13	"Excellent patency rates were achieved using both BITA configurations with no significant differences in terms of major adverse cerebro-cardiovascular events up to 19 months postoperatively or inferior temporal artery (ITA) patency."[11]	Interpretation inappropriate: power calculation incorrectly conducted on a secondary endpoint, graft patency. Our calculation of power on the primary outcome showed the study had only a 46% power to detect a 50% reduction in major adverse cerebro-cardiovascular events. The low power was not discussed as an explanation for the claim of equivalency and larger confirmatory studies were not recommended.
Transobturator tape (n = 82) compared with tension-free vaginal tape (TVT; n = 88) for the treatment of stress urinary incontinence	A composite endpoint of several parameters assessing the presence or absence of abnormal bladder function	14	"The transobturator tape is not inferior to TVT for the treatment of stress urinary incontinence and results in fewer bladder perforations. Both the objective and subjective cure rates in both groups were high, and therefore no significant differences could be detected. Larger studies are needed to evaluate the relative risk of less common but potentially severe complications."[12]	Interpretation appropriate: study appropriately identified the limitations of small sample size and recommend larger studies were recommended in the future.
Tension-free vaginal tape obturator (TVT-O; n = 132) vs tension-free vaginal tape (TVT; n = 136)	Cure rates	14	"Both the objective and subjective cure rates in both groups were high, and therefore no significant differences could be detected. Both procedures seem to be equally highly successful at 12 months postoperatively. This randomized trial shows that classic TVT and TVT-O perform equally."[13]	Interpretation inappropriate: study claimed equivalency with no discussion of power limitation and does not recommend a larger trial.
Percutaneous coronary intervention (PCI) with distal protection (I, n = 312), percutaneous coronary intervention (PCI) without distal protection (C, n = 314)	Complete (greater or equal to 70%) ST-segment resolution	6	"The routine use of distal protection by a filterwire system during primary PCI does not seem to improve microvascular perfusion, limit infarct size, or reduce the occurrence of major adverse cardiac and cerebrovascular event (MACCE). The results of the present study demonstrate that routine use of adjunctive mechanical devices cannot be advocated during PCI treatment of patients with ST-segment elevation myocardial infarction (STEMI)."[14]	Interpretation appropriate: their calculations show that 450 patients were needed for an 80% power. Study randomized 626 patients to increase power. Claim of equivalency warranted.
Cognitive outcomes after off-pump coronary artery bypass grafting (CABG; I, n = 61), on-pump CABG (C, n = 59)	Cognitive dysfunction at 12 mo after surgery	7	"There were no significant differences in the incidence of cognitive decline between the off-pump (19%) and on-pump (9%) group. The detection of a difference between 19% and 9%	Interpretation appropriate: the primary power calculation, however, was inappropriate because the rates of adverse

Study question	Primary endpoints	Secondary endpoints, n	Authors' conclusion	Our evaluation of the authors' interpretation
			would require approximately 400 patients. Hence there is a need for larger multi-center trials."[15]	outcomes used (50% and 20%) were not justified. The study does claim equivalency with low power; however, it notes the limitations of the small sample size and recommends larger studies.
CABG with intramuscular (n = 21) or intracoronary (n = 21) bone marrow cells (BMC) or standard CABG (n = 20) without intramuscular or intracoronary bone marrow cells	Improvement in systolic function of scar segments 6 mo after treatment	10	"Injection of autologous BMCs directly into the scar or into the artery supplying the scar is safe but does not improve contractility of nonviable scarred myocardium, reduce scar size, or improve left ventricular function more than CABG alone."[16]	Interpretation inappropriate: there is a trend toward a negative effect of the bone marrow treatments, yet a strong claim of equivalency is given with no discussion of the limitations of lack of power, nor were larger studies recommended.
Carotid artery stenting with (n = 18) or without (n = 18) cerebral protection.	New ischemic injury	4	"Our data suggest that distal protection filters may not be as effective as expected in reducing microemboli compared with stenting without any filter protection. New defects on MRI were noted in 13/18 (72%) of those with protective devices versus 8/18 (44%) without. Larger studies are clearly warranted."[17]	Interpretation appropriate: the study was stopped by the Data Safety Monitoring Board due to unsuccessful recruitment. The study power calculations suggested enrollment should be 45 in each group, but enrollment of only 18 in each group was achieved. Study failed to discuss the limitations of low power, however the authors did recommend larger studies.
Routine (n = 101) vs provisional T-stenting (n = 101) in the treatment of de novo coronary bifurcation lesions	Percent stenosis of the side branch at 9 mo	5	"As our key result we did not find that routine T-stenting reduced the risk of the side-branch restenosis. The per cent diameter stenosis in the side branch after provisional T-stenting was lower than projected, which reduces power to detect the projected 33% reduction in per cent diameter stenosis by routine T-stenting."[18]	Interpretation appropriate: lack of power considered and a larger study published just as the results of this study made available were discussed. Further, the authors note the high rate of stent stenosis and recommend further studies.