Acta Neurochirurgica (2019) 161:623–625
https://doi.org/10.1007/s00701-019-03850-3

EDITORIAL (BY INVITATION) - NEUROSURGERY GENERAL

Randomized controlled trials in surgery and the glass ceiling effect

Ole Solheim [1,2]

Received: 11 February 2019 / Accepted: 12 February 2019 / Published online: 23 February 2019

Richard Horton ridiculed over primitive surgical research methods in a famous commentary in the Lancet more than two decades ago [9]. Surgical research was compared with a comic opera—where questions are shouted out, but without answers. Horton stated "I should like to shame [surgeons] out of the comic opera performances which they suppose are statistics of operations." The dominating study method in neurosurgery, namely retrospective series of surgeons evaluating their own work is often hopelessly biased and the scientific value may certainly be questioned. Without well-conducted randomized controlled trials (RCTs), can we really advance and know what works and what does not? And before we discuss the optimal surgical management, do we really know that surgery helps at all?

Surgery has traditionally been more experience based than evidence based, and much of the enormous advancements in of surgical treatment over the last 200 years have taken place without rigorous scientific trials [6]. Still, there are several thought-provoking examples of surgical interventions that have failed the test of a sham-controlled randomized trial, for example arthroscopic knee surgery for degenerative osteoarthrosis or meniscal tears [15, 20], subacromial decompression for shoulder impingement [17], diaphragm pacing in ALS [7], deep brain stimulation in the ventral capsule/ventral striatum for depression [5], renal denervation for hypertension [1], and vertebroplasty for osteoporotic vertebral fractures [2]. In fact, an overwhelming majority of sham-controlled trials in surgery fail to demonstrate an effect of the intervention [18]. Does this mean that surgery often is less effective than we like to believe? Or does it merely reflect the fact that few surgical interventions can be placebo controlled and that such trials are more often done when the likelihood of surgery being an effective treatment is low in the first place? Nevertheless, very few surgical treatments have been placed under the scrutiny of such trials and many operations are still backed by relative scarce and often conflicting evidence, not at least in neurosurgery.

Martin and colleagues [14] conducted an interesting and well-written systematic review with critical appraisal of RCTs that compare neurosurgical interventions with non-operative therapy. They identified 82 neurosurgical RCTs published between 2000 and 2017. More than a third of the trials found a difference between operative and non-operative treatment, but in less than 4% non-operative management was found to be superior. However, most trials were conducted by surgeons, and this can be a source of bias as the genealogy of scientist may have an impact on their findings [8]. Further, high JADAD score, reflecting better methodological quality [10], was associated with lower chance of demonstrating a benefit from surgery. The JADAD scale places much emphasis on blinding and the inter-rater agreement for the JADAD scale has been questioned [3]. Still, almost half of the neurosurgical studies in the current review had a JADAD score of 1 or 2, indicating rather poor quality. Further, study protocols were properly registered in less than half of the trials, and among the registered trials, 13% and 34% had changed the primary or secondary endpoint between trial registration and publication. These findings are in line with previous reports [12, 13], but the results are still concerning. RCTs crown the pyramid of evidence and may change clinical practice. However, can we trust the surgical RCTs—and is RCT design really the best study design in surgery? Perhaps we are giving many surgical RCTs too much credit? Or are we simply not planning and conducting these trials well enough?

A rigorous and well-conducted RCT has several advantages. First, following a strict and published study protocol avoids post hoc (or ad hoc) result driven explorations of different endpoints and case selections. Second, the by default prospective study design ensures that assessments, documentation, and follow-up can be uniform and planned. Third,

This article is part of the Topical Collection on *Neurosurgery general*

✉ Ole Solheim
ole.solheim@ntnu.no

[1] Department of Neurosurgery, St. Olavs University Hospital, Trondheim, Norway

[2] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway

randomization or planned cross-over ensures that known or unknown prognostic factors (or confounders) are equally distributed between cases and controls. Forth, blinding ensures that potential bias associated with assessment, reporting, or analyzes can be minimalized. In the most extensive form, so-called quadruple blinding, both the patients, the study team, the outcome evaluators, and data analysts are blinded to what treatment was given. The study paper may then be written up until the discussion section before the seal is broken and the allocation is revealed. Although this rigorous form of RCTs might be feasible if comparing two white pills with similar dose repetitions and side effects, most surgical RCTs are associated with several weaknesses that are difficult to overcome. First, timing and planning of surgical RCTs are difficult since surgical methods evolve gradually over time and too early rigorous assessments may be unfair to the evolving methods which few surgeons yet master while clinical equipoise may be lost if evaluations are done too late. Also, standardization of interventions may be difficult as surgeons are trained differently, improvise, and may develop their surgical technique over time. Also, perioperative care and surveillance are often different between different departments and may be tailored in given patients. Drug trials are usually conducted in phases where endpoints, statistical power, and adverse effects are first explored in smaller phase II RCTs before the lager multicenter phase III trials are carried out. Marketing is generally not allowed before the phase III trial is finished. Conversely, surgical RCTs are usually not done in stages and choosing the most appropriate endpoints and estimations about effect sizes and statistical power can therefore be difficult. As a result, most surgical RCTs are comparable with phase II drug trials and are therefore small and sometimes more explorative in nature. This increases the likelihood of both false positive and false negative finding. After these smaller phase II-like surgical trials are done, clinical equipoise is perhaps not there anymore and validation studies and larger multicenter studies are seldom conducted. Second, recruitment may be problematic. Patients are often more reluctant towards invasive, risky, and non-reversible interventions if the treating physician has no clear treatment recommendation. Also, operative treatments are usually not as common as drug treatments, and this affects recruitment, market value, and funding opportunities. This may further contribute to underpower and false negative findings. Third, pre-inclusion bias may be problematic, not at least in studies comparing surgery with non-operative treatment. Often, patients referred to surgery have already tried several non-operative treatments. This may introduce expectation bias and reduce the likelihood of demonstrating an effect of further non-operative management. Third, cross-over is often a major problem, not at least when comparing surgery with non-operative management. For example, the SPORT study on surgical vs. non-operative treatment for lumbar disc herniation, 60% assigned to surgery had been operated at 2 years compared with 45% of patients randomized to non-operative treatment [22]. Thus, intention-to-treat analyses were much obscured and the trial ended up more observational than interventional in the end.

Perhaps the most important weakness associated with surgical trials is difficulty with blinding. In a systematic review of 250 RCTs, researchers observed considerable differences in treatment effects between trials that reported "double-blinding" compared with those that did not [19]. However, the surgeon can usually not be blinded and sham-controlled interventions are seldom feasible, especially if comparing with non-operative treatments. Thus, the great majority of such trials are open label. As the placebo (and perhaps also nocebo) effects of various interventions vary, this weakness is significant and often difficult to overcome. This is especially problematic with subjective and patient-reported endpoints and short follow-ups. Further, difference in follow-up routines or surveillance between the two treatment arms is not uncommon if comparing completely different treatment modalities. This may have large effects on its own, also on hard endpoints. For example, a recent study in lung cancer patients randomized to either continuous web-based symptom monitoring with patient reported outcomes or standard follow-up with scheduled imaging every 3 to 6 months demonstrated a large survival difference, median 22.5 months vs. 14.9 months [4].

Due to the aforementioned points, surgical RCTs usually cannot fulfill many of the important quality traits of the ideal and gold standard RCT. Consequently, the overall methodological quality is often poor, as seen in the review by Martin and colleagues [14]. According to Oxford Center for Evidence Based Medicine, the obtainable level of evidence from a poor quality RCT is not better than a well-conducted cohort study, namely level 2b (https://www.cebm.net/). At times, a good cohort study can even be more informative than a poor quality RCT. In fact, many of the important quality traits of RCTs can (and should) be copied in well-designed cohort studies. Like an RCT, a good cohort study is prospective in nature. In the lack of randomization, propensity matching can, for example be, done to limit selection bias [16]. Also, differences in favored treatments between centers may enable pseudorandomized trials [11]. Collection of prospective data, for example within treatment registries may enable parallel cohort studies or pragmatic trials with sufficient statistical power. Also, the study team, outcome evaluators, or data analysts may be blinded. Further, while the generalizability of RCTs can be limited by strict inclusion and exclusion criteria, cohort studies may capture the clinical effectiveness in everyday patients. Registration of protocols, before data is retrieved, is unfortunately still not required in observational studies, but would be a major strength to limit or expose explorative data analyses. However, while many journals embrace the STROBE statement (or checklist) for observational studies [21], these requirements only deal with transparency of reporting.

Although many surgical RCTs may reach a glass ceiling on the climb up the hierarchy of evidence, many of these trials still deserve considerable impact, not at least in comparison with the rather primitive conventional study methods in surgery. Even so, there is great potential for improving study designs in surgery, not at least for cohort studies.

References

1. Bhatt DL, Kandzari DE, O'Neill WW, D'Agostino R, Flack JM, Katzen BT, Leon MB, Liu M, Mauri L, Negoita M, Cohen SA, Oparil S, Rocha-Singh K, Townsend RR, Bakris GL (2014) A controlled trial of renal denervation for resistant hypertension. N Engl J Med 370:1393–1401. https://doi.org/10.1056/NEJMoa1402670 Epub 1402014 Mar 1402629
2. Buchbinder R, Osborne RH, Ebeling PR, Wark JD, Mitchell P, Wriedt C, Graves S, Staples MP, Murphy B (2009) A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. N Engl J Med 361:557–568. https://doi.org/10.1056/NEJMoa0900429
3. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, Laupacis A (1999) Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials 20:448–452
4. Denis F, Basch E, Septans AL, Bennouna J, Urban T, Dueck AC, Letellier C (2019) Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. Jama 321:306–307. https://doi.org/10.1001/jama.2018.18085
5. Dougherty DD, Rezai AR, Carpenter LL, Howland RH, Bhati MT, O'Reardon JP, Eskandar EN, Baltuch GH, Machado AD, Kondziolka D, Cusin C, Evans KC, Price LH, Jacobs K, Pandya M, Denko T, Tyrka AR, Brelje T, Deckersbach T, Kubu C, Malone DA Jr (2015) A randomized sham-controlled trial of deep brain stimulation of the ventral capsule/ventral striatum for chronic treatment-resistant depression. Biol Psychiatry 78:240–248. https://doi.org/10.1016/j.biopsych.2014.1011.1023 Epub 2014 Dec 1013
6. Gawande A (2012) Two hundred years of surgery. N Engl J Med 366:1716–1723. https://doi.org/10.1056/NEJMra1202392
7. Gonzalez-Bermejo J, Morelot-Panzini C, Tanguy ML, Meininger V, Pradat PF, Lenglet T, Bruneteau G, Forestier NL, Couratier P, Guy N, Desnuelle C, Prigent H, Perrin C, Attali V, Fargeot C, Nierat MC, Royer C, Menegaux F, Salachas F, Similowski T (2016) Early diaphragm pacing in patients with amyotrophic lateral sclerosis (RespiStimALS): a randomised controlled triple-blind trial. Lancet Neurol 15:1217–1227. https://doi.org/10.1016/S1474-4422(1216)30233-30232 Epub 32016 Oct 30211
8. Hirshman BR, Alattar AA, Dhawan S, Carley KM, Chen CC (2019) Association between medical academic genealogy and publication outcome: impact of unconscious bias on scientific objectivity. Acta Neurochir 23:019–03804
9. Horton R (1996) Surgical research or comic opera: questions, but few answers. Lancet. 347:984–985
10. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ (1996) Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 17:1–12
11. Jakola AS, Skjulsvik AJ, Myrmel KS, Sjavik K, Unsgard G, Torp SH, Aaberg K, Berg T, Dai HY, Johnsen K, Kloster R, Solheim O (2017) Surgical resection versus watchful waiting in low-grade gliomas. Ann Oncol 28:1942–1948. https://doi.org/10.1093/annonc/mdx1230
12. Kiehna EN, Starke RM, Pouratian N, Dumont AS (2011) Standards for reporting randomized controlled trials in neurosurgery. J Neurosurg 114:280–285. https://doi.org/10.3171/2010.3178.JNS091770 Epub 092010 Nov 091775
13. Mansouri A, Cooper B, Shin SM, Kondziolka D (2016) Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered? J Neurosurg 124:558–568. https://doi.org/10.3171/2014.3112.JNS142465 Epub 142015 Aug 142428
14. Martin E, Muskens IS, Senders JT, DiRisio AC, Karhade AV, Zaidi HA, Moojen WA, Peul WC, Smith TR, Broekman MLD (2019) Randomized controlled trials comparing surgery to non-operative management in neurosurgery: a systematic review. Acta Neurochir. https://doi.org/10.1007/s00701-019-03849-w
15. Moseley JB, O'Malley K, Petersen NJ, Menke TJ, Brody BA, Kuykendall DH, Hollingsworth JC, Ashton CM, Wray NP (2002) A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med 347:81–88. https://doi.org/10.1056/NEJMoa013259
16. Nerland US, Jakola AS, Solheim O, Weber C, Rao V, Lonne G, Solberg TK, Salvesen O, Carlsen SM, Nygaard OP, Gulati S (2015) Minimally invasive decompression versus open laminectomy for central stenosis of the lumbar spine: pragmatic comparative effectiveness study. Bmj. 350:h1603. https://doi.org/10.1136/bmj.h1603
17. Paavola M, Malmivaara A, Taimela S, Kanto K, Inkinen J, Kalske J, Sinisaari I, Savolainen V, Ranstam J, Jarvinen TLN (2018) Subacromial decompression versus diagnostic arthroscopy for shoulder impingement: randomised, placebo surgery controlled clinical trial. Bmj. 362:k2860. https://doi.org/10.1136/bmj.k2860
18. Probst P, Grummich K, Harnoss JC, Huttner FJ, Jensen K, Braun S, Kieser M, Ulrich A, Buchler MW, Diener MK (2016) Placebo-controlled trials in surgery: a systematic review and meta-analysis. Medicine (Baltimore) 95:e3516. https://doi.org/10.1097/MD.0000000000003516
19. Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Jama. 273:408–412
20. Sihvonen R, Paavola M, Malmivaara A, Itala A, Joukainen A, Nurmi H, Kalske J, Jarvinen TL (2013) Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. N Engl J Med 369:2515–2524. https://doi.org/10.1056/NEJMoa1305189
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370:1453–1457. https://doi.org/10.1016/S0140-6736(1407)61602-X
22. Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, Abdu WA, Hilibrand AS, Boden SD, Deyo RA (2006) Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. Jama 296:2441–2450. https://doi.org/10.1001/jama.2296.2420.2441