North American Spine Society Journal (NASSJ) 3 (2020) 100019

Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

journal homepage: www.elsevier.com/locate/xnsj

Editorial

Evidence-based medicine and clinical decision-making in spine surgery

The adoption of evidence-based medicine's (EBM) principles for medical research has been one of the greatest scientific breakthroughs of the twentieth century [1,2]. In fact, for scientists and physicians of our generation, to whom systematic reviews, meta-analyses and clinical guidelines are an essential part of our scientific landscape, it seems hard to believe that before the passage of the U.S. Kefauver-Harris Amendment in 1962, testing of new drugs and medical devices in human clinical trials was not even a legal requirement for obtaining approval by the Food and Drug Administration (FDA) [3].

However, unlike some of the classic trials in medical specialties, which have led to major advances [4–6], the history of clinical trials in surgical specialties has been somewhat less grandiose, in part due to the natural challenges involved in randomization and blinding of surgical patients [7,8]. This is certainly true for spine surgery. In this editorial, some of these clinical trials will be discussed with the goal of establishing a few heuristic principles on how to properly evaluate the practical implications of EBM results while avoiding uncritical and blind reliance on "high-quality clinical evidence".

NASCIS-2: the perils of multiple subgroup analyses

The Second National Acute Spinal Cord Injury Study (NASCIS-2) was a prospective randomized clinical trial which evaluated the outcomes of high-dose methylprednisolone (bolus of 30 mg/kg, followed by a continuous infusion of 5.4 mg/kg/h for 23 h) versus placebo for patients with acute spinal cord injury (SCI) presenting within 12 h of the initial traumatic event [9]. Although at 1-year follow-up there were no differences in neurological outcomes between both groups, a subgroup analysis suggested that patients who received steroids within 8 h had superior outcomes in terms of both sensory and motor function motor at 6 months.

Among other criticisms [10], it has been pointed that stratification based on an 8-hour timeframe was not part of the initial design and, therefore, data dredging (also called p-hacking) through multiple subgroup analyses using different timeframes and subcategories may have led to possible spurious findings. It has been estimated that, by subdividing patients in complete and incomplete injuries, paraplegic, tetraplegic and paretic patients, among several other groupings, at least 27 subgroup analyses were performed with the obtained data.

As it has been classically demonstrated by an interesting subgroup analysis included in the original manuscript of the Second International Study of Infarct Survival (ISIS-2) [11], even a bizarre stratification of patients according to astrological signs may be enough to change the status of statistical significance between the intervention and the control groups. Finally, it should be highlighted that in NASCIS-2, the placebo group treated within 8 h did worse not only when compared with the methylprednisolone group treated within 8 h but also when compared with the placebo group treated after 8 h, possibly suggesting a significant imbalance between such groups at baseline [12].

Although it actually took more than 2 decades before guidelines from professional organizations, including the Congress of Neurological Surgeons (CNS) and the American Association of Neurological Surgeons (AANS), published formal recommendations against the use of high-dose methylprednisolone therapy in patients with acute SCI [13], there were in fact some early criticisms regarding the way such a trial was conducted, presented and interpreted [14]. During this period a whole generation of spine surgeons has routinely prescribed high-dose methylprednisolone for the treatment of acute spinal cord injury, with a significant proportion of physicians doing so mainly because of fear of litigation [15], despite the fact that there has never been formal FDA approval of methylprednisolone for such an indication.

2016. NEJM trials on spinal fusion: the issue of generalizability

Another interesting exercise on how to properly interpret the results of clinical studies in spine surgery involves two prospective randomized trials which were published in the same volume of the New England Journal of Medicine (NEJM) in 2016 [16,17].

The first one, known as the Swedish Spinal Stenosis Study, randomized patients with spinal stenosis with or without degenerative spondylolisthesis to decompression alone or decompression with fusion. The study demonstrated no statistical difference in the Oswestry Disability Index (ODI) or in the 6-minute walk test between both groups at the 2 and 5-years follow-up, although as expected, operative time, intraoperative blood loss and costs were higher in the fusion group. Based on such results the authors claimed that, among patients with lumbar stenosis with or without spondylolisthesis, the addition of fusion had no substantial benefit in terms of long-term outcomes [16].

The other study published by several well-known spine surgeons in North America, randomized patients with stable grade 1 spondylolisthesis and associated lumbar canal stenosis to decompression alone or decompression and fusion. The study demonstrated a greater increase in the SF-36 physical-component summary (PCS) scores in the surgical group at the 2-year follow-up which persisted at the 3 and 4-years follow-up, although no differences were observed in the ODI. The cumulative rate of re-operation was also different between both groups (34% in the non-instrumented group and 14% in the instrumented group - $P = 0.05$). Based on such results the authors argued that for patients with stable grade 1 spondylolisthesis, decompression with instrumented fusion had a slightly greater but clinically meaningful impact upon long-

https://doi.org/10.1016/j.xnsj.2020.100019
Received 28 July 2020; Received in revised form 2 August 2020; Accepted 2 August 2020
Available online 5 August 2020
2666-5484/© 2020 The Author. Published by Elsevier Ltd on behalf of North American Spine Society. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

term physical health–related qualify of life outcomes as well as lower re-operation rates when compared to decompression alone [17].

Although there are multiple ways to try to reconcile the apparently contradictory results of these two studies in terms of the clinical efficacy of lumbar fusion, a few remarks are pertinent. In the Swedish study no flexion–extension x-rays for evaluation of segmental instability were obtained pre-operatively, which is a significant difference from the standard practice adopted by the vast majority of spine surgeons. According to the study's supplementary appendix, although 90% of the fusion procedures were instrumented posterolateral fusions, only 6 cases were submitted to interbody fusion. It should be noted that, at least in North America, a substantial proportion of instrumented lumbar fusion procedures involves an interbody cage (through either TLIF, ALIF or XLIF/DLIF/OLIF) [18], techniques which have been associated with higher fusion rates and greater restoration of foraminal height and segmental lordosis.

Additionally, a significant proportion of such procedures are performed through a minimally-invasive approach, which has been suggested to be associated with decreased perioperative blood loss and hospital stay, less tissue damage to the paraspinal muscles and possibly superior long-term functional outcomes, especially regarding back pain, when compared to open procedures [19]. Therefore, it could be reasonably argued that all the Swedish study demonstrated is that if patients with lumbar stenosis are selected for fusion without a standard protocol for investigation of spinal instability and are operated with old techniques without interbody fusion or minimally invasive approaches, the results of such poorly indicated (and possibly sub-optimally performed) fusions are no different than those of decompression alone. Conversely the North American study demonstrated that, even excluding patients with documented instability (which are the ones who would likely benefit the most from a fusion) and considering only patients with stable grade 1 spondylolisthesis, it seems that instrumented fusion in addition to decompression is associated with lower rates of re-operation and somewhat superior long-term outcomes in terms of quality of life.

It should be noted that the authors' claim about a "slightly greater but clinically meaningful improvement in overall physical health-related quality of life" is debatable, especially as other studies have demonstrated the minimal clinically important difference (MCID) for SF-36-PCS to be higher (4.9 according to Rampersaud et al. [20] and 10 according to Adogwa et al. [21], both at 2 year follow-up) than the 3.2 difference observed in this study. I am confident other interpretations of these two studies are plausible and possibly even persuasive. The important point to be highlighted here is that, quite often, different high-quality studies according to EBM standards will demonstrate apparently paradoxical results which require a thoughtful and critical analysis of each study's design, conduction and conclusions before such results can be properly translated to the daily clinical practice.

SPORT: challenges with randomization and intention-to-treat analysis

Another clinical study in spine surgery which provides a few interesting lessons is the Spine Patient Outcomes Research Trial (SPORT) trial, a large $13.5 million NIH-funded study which, among other lumbar spine pathologies, compared outcomes of surgery versus conservative treatment for patients with symptomatic lumbar disk herniation [22]. Although the observational SPORT disk herniation cohort study suggested superiority of surgery over conservative treatment [23], the randomized trial failed to demonstrate a statistically significant difference between the operative and non-operative arms at all time-points.

The failure of SPORT to demonstrate a statistically significant difference between both groups seemed to be largely related to the very high cross-over rates (at 3 months only 50% of patients assigned to the operative group actually received surgery, while 30% of those assigned to non-operative treatment received surgery in the same period), which substantially undermined the results of the intention-to-treat analysis. As previously highlighted [24], instead of finally demonstrating through EBM standards the efficacy of one of the most commonly performed and well-established procedures in spine surgery, all the SPORT study was able to show was that, regardless of randomization attempts, patients with severe pain will ultimately undergo surgery and present good long-term outcomes while those with mild symptoms will choose to continue conservative treatment with comparable long-term outcomes.

STASCIS: what degree of evidence is enough?

The Surgical Timing In Acute Spinal Cord Injury Study (STASCIS) stimulates another important discussion about clinical studies in spine surgery, namely, the necessary level of evidence which should be required before a certain therapy can be recommended [25].

Several methodological criticisms have been raised regarding STASCIS [26], such as the absence of a proper power analysis, absence of randomization, use of methylprednisolone and hypertensive therapy at the discretion of the treating physician, baseline discrepancies in demographics and neurological function between early and late surgery groups as well as a high heterogeneity in terms of both the selected surgical approach and the type of spinal cord/spinal column injury.

Despite such factors, which raise real questions of how confident one can be about the superiority of early versus late surgical intervention for treatment of acute SCI, it should be highlighted that, most importantly, the study demonstrated no difference in medical or surgical complications as well as death between both groups. Admittedly STASCIS provides at best level 2 evidence supporting the advantages of early surgery for SCI. However, in face of the extensive literature showing the importance of the secondary injury cascade in the pathophysiology of SCI [27,28] as well as other cohort studies suggesting similar benefits of early decompression [29–31], would it not be fair to question if, taking into account the absence of deleterious effects, for patients with incomplete spinal cord injury the most appropriate conduct at this point would be to strongly consider early surgery unless prohibitive from the medical standpoint?

Such type of situation illustrates one of the most important points when considering the level of evidence for the treatment of spinal pathologies. As the current status of scientific evidence can only carry us so far in so many subjects in spine surgery and, as absence of evidence does not necessarily mean evidence of absence, would it be unreasonable to consider the default mode a certain intervention whenever it has been proven to be as safe as (even if not definitively superior) to the traditional treatment approach? While pursuing the highest level of scientific evidence on the issue of timing of surgical decompression for acute SCI, isn't the best available evidence so far, as summarized by systematic reviews and meta-analyses [32], enough to support a recommendation for early intervention in patients with acute SCI whenever feasible?

These are essentially philosophical/non-scientific questions which exemplify how, ultimately, the final responsibility for a sensible and thoughtful decision-making based on the best available evidence is on the shoulders of each treating physician. In other words, it is fine (and actually highly recommended) to pursue and rely upon high-quality scientific evidence for daily decision-making in spine surgery, but this by no means exempt us from the inherent responsibility of employing our best clinical judgement based on a critical and individualized risk-benefit analysis of available treatment options for each patient. The peril is not to rely on EBM standards, but to automatically abrogate basic principles of critical decision-making just because the evidence may be inconclusive or pointing otherwise [33].

The reproducibility crisis in science and the P-value debate

It should be noted that, these type of challenging methodological questions about the validity and generalizability of currently available

research data, pervade the scientific enterprise as a whole. Despite several warnings about the crisis of reproducibility in medical research [34] as well as calls for going beyond a simplistic application of the so-called null hypothesis significance testing paradigm [35], the vast majority of scientific research in spine surgery still relies on a dichotomous interpretation of results based on a pre-specified p-value threshold. Such automatic reliance on a specific p-value for determining the statistical significance as well as possible clinical impact of a certain therapy becomes even more problematic if considering the fact that, as pointed by expert statisticians, the difference between statistically significant and non-statistically significant is itself not statistically significant [36]. It has been shownt that simplistic solutions, such as lowering the p-value to 0.005 [37], although clearly decreasing the rates of false positives, might have the undesirable practical effect of reducing even more the availability of high quality of scientific evidence in surgical specialties such as neurosurgery [38]. In this regard, the use of confidence intervals and effect sizes for proper estimation of the magnitude of an observed effect as well as other available statistical techniques, especially those relying on a Bayesian approach as a complement to traditional frequentist analyses, should be strongly considered [39,40].

Conclusions

Despite the inherent limitations associated with its cursory nature, the present analysis provides some important lessons about the quest for scientific evidence in spine surgery. Prospective randomized clinical trials in spine surgery are not only challenging in terms of their design and conduction (as demonstrated by SPORT) but also in terms of their proper interpretation (as revealed by NASCIS-2). Even more challenging seems to be the decision of how to interpret low-quality of evidence in pathologies associated with high morbidity rates, as illustrated by STASCIS. Finally, as revealed by the 2016 NEJM trials on spinal fusion, there is also no lack of apparently paradoxical results between high-quality studies.

Although we do have the privilege of having high level evidence in a few important subjects in spine surgery, especially when considering new spinal devices [41], we are undeniably in the very early stages of our quest for high-quality scientific evidence for most of our routine practices in spine surgery. May we always remember that equally, or perhaps even more important than our final goals in such a laudable scientific endeavor may be the quality of the process through our daily decisions are made while the evidence is not there yet.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.xnsj.2020.100019.

Tobias A. Mattei
Division of Neurological Surgery, Saint Louis University - Saint Louis/MO, SSM Saint Louis University Hospital, 1008 S. Spring Avenue, St. Louis, MO 63110, United States

E-mail address: tobias.mattei@health.slu.edu
Revised 2 August 2020

References

[1] Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet 2017;390(10092):415–23. doi:10.1016/S0140-6736(16)31592-6.
[2] Chung KC, Ram AN. Evidence-based medicine: the fourth revolution in American medicine? Plast Reconstr Surg 2009;123(1):389–98. doi:10.1097/PRS.0b013e3181934742.
[3] Peltzman Sam. An evaluation of consumer protection legislation: the 1962 drug amendments. J Polit Econ Sep.-Oct., 1973;81(5):1051.
[4] Czeizel AE, Dudás I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. N Engl J Med 1992;327(26):1832–5. doi:10.1056/NEJM199212243272602.
[5] MERIT-HF Study GroupEffect of metoprolol CR/XL in chronic heart failure: metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). Lancet 1999;353(9169):2001–7.
[6] Liggins GC, Howie RN. A controlled trial of antepartum glucocorticoid treatment for prevention of the respiratory distress syndrome in premature infants. Pediatrics 1972;50:515.
[7] Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. Trials 2009;10:9 Published 2009 Feb 6. doi:. doi:10.1186/1745-6215-10-9.
[8] Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. Lancet 2009;374(9695):1097–104. doi:10.1016/S0140-6736(09)61086-2.
[9] Bracken MB, Shepard MJ, Collins WF, et al. A randomized, controlled trial of methylprednisolone or naloxone in the treatment of acute spinal-cord injury. Results of the second national acute spinal cord injury study. N Engl J Med 1990;322(20):1405–11. doi:10.1056/NEJM199005173222001.
[10] Nesathurai S. Steroids and spinal cord injury: revisiting the NASCIS 2 and NASCIS 3 trials. J Trauma 1998;45(6):1088–93. doi:10.1097/00005373-199812000-00021.
[11] ISIS-2 Collaborative GroupRandomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (second international study of infarct survival) collaborative group. Lancet 1988;2(8607):349–60.
[12] Coleman WP, Benzel D, Cahill DW, et al. A critical appraisal of the reporting of the national acute spinal cord injury studies (II and III) of methylprednisolone in acute spinal cord injury. J Spinal Disord 2000;13(3):185–99. doi:10.1097/00002517-200006000-00001.
[13] Hurlbert RJ, Hadley MN, Walters BC, et al. Pharmacological therapy for acute spinal cord injury. Neurosurgery 2013;72(suppl 2):93–105 doi:10.1227.
[14] Hanigan WC, Anderson RJ. Commentary on NASCIS-2. J Spinal Disord 1992;5(1):125–33. doi:10.1097/00002517-199203000-00019.
[15] Falavigna A, Quadros FW, Teles AR, et al. Worldwide steroid prescription for acute spinal cord injury. Global Spine J 2018;8(3):303–10. doi:10.1177/2192568217735804.
[16] Försth P, Ólafsson G, Carlsson T, et al. A randomized, controlled trial of fusion surgery for lumbar spinal stenosis. N Engl J Med 2016;374(15):1413–23. doi:10.1056/NEJMoa1513721.
[17] Ghogawala Z, Dziura J, Butler WE, et al. Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis. N Engl J Med 2016;374(15):1424–34. doi:10.1056/NEJMoa1508788.
[18] Saifi C, Cazzulino A, Laratta J, et al. Utilization and economic impact of posterolateral fusion and posterior/transforaminal lumbar interbody fusion surgeries in the United States. Global Spine J 2019;9(2):185–90. doi:10.1177/2192568218790557.
[19] Qin R, Liu B, Zhou P, et al. Minimally invasive versus traditional open transforaminal lumbar interbody fusion for the treatment of single-level spondylolisthesis grades 1 and 2: a systematic review and meta-analysis. World Neurosurg 2019;122:180–9. doi:10.1016/j.wneu.2018.10.202.
[20] Rampersaud YR, Fisher C, Yee A, et al. Health-related quality of life following decompression compared to decompression and fusion for degenerative lumbar spondylolisthesis: a Canadian multicentre study. Can J Surg 2014;57(4):E126–33. doi:10.1503/cjs.032213.
[21] Adogwa O, Elsamadicy AA, Han JL, Cheng J, Karikari I, Bagley CA. Do measures of surgical effectiveness at 1 year after lumbar spine surgery accurately predict 2-year outcomes? J Neurosurg Spine 2016;25(6):689–96. doi:10.3171/2015.8.SPINE15476.
[22] Weinstein JN, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the spine patient outcomes research trial (SPORT): a randomized trial. JAMA 2006;296(20):2441–50. doi:10.1001/jama.296.20.2441.
[23] Weinstein JN, Lurie JD, Tosteson TD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. JAMA 2006;296(20):2451–9. doi:10.1001/jama.296.20.2451.
[24] Angevine PD, McCormick PC. SPORT: what neurosurgeons need to know. Clin Neurosurg 2008;55:72–5.
[25] Fehlings MG, Vaccaro A, Wilson JR, et al. Early versus delayed decompression for traumatic cervical spinal cord injury: results of the surgical timing in acute spinal cord injury study (STASCIS). PLoS ONE 2012;7(2):e32037. doi:10.1371/journal.pone.0032037.
[26] O'Toole JE. Timing of surgery after cervical spinal cord injury. World Neurosurg 2014;82:e389–90.
[27] Oyinbo CA. Secondary injury mechanisms in traumatic spinal cord injury: a nugget of this multiply cascade. Acta Neurobiol Exp (Wars) 2011;71(2):281–99.
[28] Choo AM, Liu J, Dvorak M, Tetzlaff W, Oxland TR. Secondary pathology following contusion, dislocation, and distraction spinal cord injuries. Exp Neurol 2008;212(2):490–506. doi:10.1016/j.expneurol.2008.04.038.
[29] Jug M, Kejžar N, Vesel M, et al. Neurological recovery after traumatic cervical spinal cord injury is superior if surgical decompression and instrumented fusion are performed within 8 h versus 8 to 24 h after injury: a single center experience. J Neurotrauma 2015;32(18):1385–92.
[30] Wilson JR, Singh A, Craven C, et al. Early versus late surgery for traumatic spinal cord injury: the results of a prospective Canadian cohort study. Spinal Cord 2012;50(11):840–3. doi:10.1038/sc.2012.59.
[31] Mattiassich G, Gollwitzer M, Gaderer F, et al. Functional outcomes in individuals undergoing very early (< 5 h) and early (5-24 h) surgical decompression in traumatic cervical spinal cord injury: analysis of neurological improvement from the Austrian spinal cord injury study. J Neurotrauma 2017;34(24):3362–71. doi:10.1089/neu.2017.5132.
[32] Liu JM, Long XH, Zhou Y, Peng HW, Liu ZL, Huang SH. Is urgent decompression superior to delayed surgery for traumatic spinal cord injury? A meta-analysis. World Neurosurg 2016;87:124–31. doi:10.1016/j.wneu.2015.11.098.

[33] Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 2003;327(7429):1459–61. doi:10.1136/bmj.327.7429.1459.
[34] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124. doi:10.1371/journal.pmed.0020124.
[35] Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biol Lett 2019;15(5):20190174. doi:10.1098/rsbl.2019.0174.
[36] Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat 2006;60:328–31.
[37] Ioannidis JPA. The proposal to lower P value thresholds to .005. JAMA 2018;319:1429–30.
[38] Mattei TA. Practical effects of lowering the P value in neurosurgery: restricting evidence-based medicine to big business. World Neurosurg 2018;117:460–2. doi:10.1016/j.wneu.2018.07.112.
[39] Lin R, Yin G. Bayes factor and posterior probability: complementary statistical evidence to p-value. Contemp Clin Trials 2015;44:33–5. doi:10.1016/j.cct.2015.07.001.
[40] Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology 2013;24(1):62–8. doi:10.1097/EDE.0b013e3182785741.
[41] Cher DJ, Capobianco RA. Spine device clinical trials: design and sponsorship. Spine J 2015;15(5):1133–40. doi:10.1016/j.spinee.2015.01.027.